Bayesian Analysis (0000) **00**, Number 0, pp. 1

Optimal Shrinkage Estimation of Predictive Densities under α–divergences

Edward George[*], Gourab Mukherjee[†] and Keisuke Yano[‡]

Abstract. We consider the problem of estimating the predictive density in a heteroskedastic Gaussian model under general divergence loss. Based on a conjugate hierarchical set-up, we consider generic classes of shrinkage predictive densities that are governed by location and scale hyper-parameters. For any α-divergence loss, we propose a risk-estimation based methodology for tuning these shrinkage hyper-parameters. Our proposed predictive density estimators enjoy optimal asymptotic risk properties that are in concordance with the optimal shrinkage calibration point estimation results established by Xie, Kou, and Brown [53] for heteroskedastic hierarchical models. These α-divergence risk optimality properties of our proposed predictors are not shared by empirical Bayes predictive density estimators that are calibrated by traditional methods such as maximum likelihood and method of moments. We conduct several numerical studies to compare the non-asymptotic performance of our proposed predictive density estimators with other competing methods and obtain encouraging results.

MSC2020 subject classifications: Primary 62L20; secondary 60F15, 60G42.

Keywords: predictive density, α-divergences, predictive inference, optimal shrinkage, risk estimation, empirical Bayes.

1 Introduction

Predictive density estimation (prde) is one of the fundamental problems in statistical prediction analysis (see chapters 2, 7 and 10 of Aitchison and Dunsmore [1] and chapters 2, 3 and 9 of Geisser [14]). Predictive density estimates assign probabilities to all possible future outcomes and can be used for better risk assessment and decision making than traditional point estimation methods [38, 32, 54]. Predictive densities have been widely used in a host of statistical applications in weather forecasting [49], finance [48], information theory [33, 59, 3] as well as for model diagnostics and validation [15, 42, 16].

In this paper, we consider multivariate predictive density estimation under general divergence loss in a heteroskedastic Gaussian model. For point estimation, since the seminal work of James and Stein [26] there has been substantial research toward understanding the risk properties of shrinkage estimators for the homoscedastic hierarchical normal models (see Fourdrinier et al. [13], Efron [10] and the references therein). The concept of shrinkage is important because it provides an elegant framework for combining information from related populations and often leads to substantial improvements in the performances of estimators used for simultaneous inference. Komaki

[*] Department of Statistics, University of Pennsylvania, edgeorge@wharton.upenn.edu
[†] Department of Data Sciences and Operations, University of Southern California, gourab@usc.edu
[‡] The Institute of Statistical Mathematics, yano@ism.ac.jp

DOI: 0000

[29, 30], George et al. [17], Brown et al. [6] demonstrated the critical role of shrinkage priors for constructing efficient predictive density estimates (prdes) under Kullback-Leibler (KL) loss. High-dimensional decision theoretic parallels between prde under Kullback-Leibler loss and point-estimation under quadratic loss have been established in [12, 55, 18, 31, 40, 39, 58, 20]. Ghosh et al. [21], Suzuki and Komaki [45], Maruyama and Strawderman [37], L'Moudden and Marchand [35] and Ghosh and Kubokawa [19] extended those parallels for prde under general α-divergences. Using KL loss as a divergence measure between the true and estimated predictive density leads to convenient tractable analysis. However, predictive densities calibrated by KL loss are often non-robust to outliers and may under-estimate the variance or ignore important local attributes of the true density. To circumvent these issues, it is becoming increasingly popular in complex prediction approaches [50, 24, 7, 61] to use the class of α-divergences [2, 34, 4] that covers a wide spectrum of divergence measures with contrasting attributes.

For predictive density estimation in multivariate Gaussian models, Ghosh et al. [21] showed that the canonical minimax prde, which is the Bayes prde under uniform prior, is not admissible under general divergence loss for dimensions greater than 2. For dominating the canonical minimax prde, Ghosh et al. [21] used prdes that are not necessarily Bayes, whereas Maruyama et al. [36] established the domination results for the Bayes prde under the harmonic prior of [17]. Ghosh and Kubokawa [19] established that the hierarchical Bayes prde has lower frequentist risk than that of the empirical Bayes prde in a regression set-up. While [21, 19] showcased enhanced predictive efficiency of the Bayes prde from non-informative priors over plug-in prdes, L'Moudden and Marchand [35] proposed improving plug-in prdes directly. However, all of these results are based on the homoskedastic model. They also do not provide any prescription for selecting a particular prde among the host of feasible and admissible prdes.

Here, we study the prde in a heteroskedastic set-up where our target density is no longer spherically symmetric, and consider the problem of finding optimal shrinkage directions. We provide a data driven program for determining the optimal directions (location) and magnitude (scale) of shrinkage such that the resultant prde has minimal frequentist risk among a wide class of shrinkage estimators. Our proposed prde not only possesses the plug-in-dominance properties of the Bayes prde as in [21], but also obtains the minimal risk among a wide class of shrinkage rules. These α-predictive risk optimality properties parallel those established by Xie et al. [53] for point estimation in heteroskedastic hierarchical models.

Recent point estimation results of Xie et al. [53, 52], Tan [47], Weinstein et al. [51] have brought to light new shrinkage phenomena in heteroskedastic models. A hierarchical set-up specifying a second-level structure to motivate the shrinkage is considered, and the corresponding hyper-parameters are subsequently estimated. Whereas, the common practice is to choose the conjugate hierarchical structure and estimate the hyper-parameters through empirical Bayes maximum likelihood estimator (EBMLE) or empirical Bayes method of moments (EBMOM), we instead consider tuning the hyper-parameters by minimizing efficient risk estimates as in [53].

A significant finding reported in [53, 52, 47, 51] is that, under heteroskedasticity, EBMLE or EBMOM provide sub-optimal predictive performance and are far outperformed by algorithms tuned using risk estimation-based approaches. We establish asymptotic optimality of our proposed predictive methods akin to the point estimation results in Xie et al. [53]. These asymptotic properties are not shared by EBMLE or EBMOM based prdes. We establish asymptotic convergence rates of our risk estimates as dimension increases. Dimension independent non-asymptotic characterizations of the predictive risk of our proposed estimators is also provided using maximal inequalities for martingales. We compare these comprehensive results for general α-divergence with those of Xu and Zhou [56] who studied empirical Bayes prde in spherically symmetric homoskedastic Gaussian model under KL loss. Our general α-divergence results well reconcile with the KL results in the existing literature. Through numerical studies, we demonstrate the benefits of using α-divergence based risk calibrated prdes over EBMLE or EBMOM based prdes. The direction of shrinkage and the shape of the optimally shrunken prdes greatly varies as α changes.

2 Predictive Set-up

Predictive Sequence Model. Consider observing a vector $\boldsymbol{X} = \{X_1, \dots, X_n\}$ where X_i are independent among each other and X_i follows $N(\theta_i, \sigma_i^2)$, $i = 1, \dots, n$. Based on observing $\boldsymbol{X}$ we would like to predict the unknown density of future observations $\boldsymbol{Y} = \{Y_i : 1 \le i \le n\}$, where Y_i independently follow $N(\theta_i, \nu_i^2)$. Here, σ_i, ν_i are known and thus, $r_i = \nu_i^2/\sigma_i^2$ is the known ratio of the future-to-past variances. The observed past $\boldsymbol{X}$ and unobserved future $\boldsymbol{Y}$ are only related through the unknown location parameter $\boldsymbol{\theta} = \{\theta_i : 1 \le i \le n\}$. This is the heteroskedastic version of the Gaussian predictive model studied in [29, 17, 6, 56]. Let $\hat{p}(\boldsymbol{y}|\boldsymbol{x})$ be any prde for the true density $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\nu}) = \prod_{i=1}^n \phi(y_i - \theta_i; \nu_i)$ of $\boldsymbol{Y}$. Note that, here we denote normal probability density function (pdf) with variance v by $\phi(\cdot\,; v)$ and thus, $\phi(y_i - \theta_i; \nu_i) = \nu_i^{-1/2}\phi(\nu^{-1/2}(y_i - \theta_i))$ where $\phi(\cdot)$ denotes standard normal pdf.

Loss Function. Consider α-divergence as the measure of discrepancy between the prde and the true density. For any fixed $\alpha \in \mathbb{R}$, the loss defined as

$$L_{n,\alpha}(\boldsymbol{\theta}, \hat{p}(\,\cdot\,; \boldsymbol{x})) := \int p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\nu})\, \ell_\alpha\left(\frac{\hat{p}(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\nu})}\right) d\boldsymbol{y},$$

where the *f-divergence* function ℓ_α with domain in $[0, \infty)$ is:

$$\ell_\alpha(z) := \begin{cases} \{4/(1-\alpha^2)\}\{1 - z^{(1+\alpha)/2}\}, & \alpha \neq \pm 1, \\ -\log z, & \alpha = -1, \\ z \log z, & \alpha = 1. \end{cases}$$

Integrating over the density of the observed past, we have the predictive α-risk as:

$$R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}) := \int p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\sigma}) L_{n,\alpha}(\boldsymbol{\theta}, \hat{p}(\,\cdot\,; \boldsymbol{x}))\, d\boldsymbol{x}.$$

When $\alpha = 3$, we have Pearson χ^2-divergence where as $\alpha = -3$ corresponds to Neyman χ^2-divergence; $\alpha = -1$ is KL loss and $\alpha = 1$ yields reverse KL loss; $\alpha = 0$

corresponds to Bhattacharyya-Hellinger (BH) loss [5]. Note that, the predictive risk might not be well-defined for all $\alpha \in \mathbb{R}$ on which we reflect later. For $\alpha_0 \in \{-1, 1\}$, $R_{n,\alpha_0}(\boldsymbol{\theta}, \hat{p}) = \lim_{\alpha \to \alpha_0} R_{n,\alpha}(\boldsymbol{\theta}, \hat{p})$, so that the KL and reverse KL predictive risk expressions will follow from the general α–predictive risk. Note that, the α predictive risk is the posterior predictive relative entropy regret criterion introduced in [46] and differs from the reference prior inducing criterion studied in [8, 22].

Hierarchical set-up and Shrinkage prdes. Next, we assume a hierarchical higher level exchangeable structure on the unknown location parameters. Let $\{\theta_i : 1 \le i \le n\}$ be independent and identically distributed (i.i.d.) from a $N(\eta, \tau)$ prior where $\eta \in \mathbb{R}$ and $\tau \ge 0$ are unknown hyper-parameters. This exchangeable hierarchical set-up, well-used in the literature [53, 11, 43, 60], allows partial pooling of information from quantities of interest for different yet related groups of populations. Define the integrated Bayes risk with respect to any prior π_n on $\boldsymbol{\theta}$ as $B_n(\pi, \hat{p}) = \int R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}) \pi_n(\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Let $\alpha_U = 1 + 2\min\{r_i : 1 \le i \le n\}$, $b = (1-\alpha)/2$ and $\bar{b} = 1 - b$. The following result (proved in section 7.2) shows that for all $\alpha \le \alpha_U$ and for any product normal priors, i.e., $\pi_n(\boldsymbol{\theta}) = \prod \phi(\theta_i - \eta; \tau)$, the integrated Bayes risk has a well-defined minima and the resultant Bayes predictive density estimator is also a product of normal densities.

Lemma 2.1. *Consider $\theta_1, \ldots, \theta_n \overset{i.i.d.}{\sim} N(\eta, \tau)$ where $\eta \in \mathbb{R}$ and $\tau \ge 0$. Then, the Bayes predictive density estimate with respect to α-divergence loss for any fixed $\alpha \le \alpha_U$ is*

$$\hat{p}[\eta, \tau](\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} \phi\big(y_i - (\omega_i x_i + \bar{\omega}_i \eta);\ (r_i + b\omega_i)\sigma_i^2\,\big) \tag{2.1}$$

where $\omega_i = \tau/(\tau + \sigma_i^2)$ and $\bar{\omega}_i = 1 - \omega_i$.

Henceforth, we consider only α–divergences with $\alpha \le \alpha_U$. Note that as $\min_i r_i \ge 0$, BH, KL, reverse KL, Neyman χ^2 divergences are always covered in our results. If $\min_i r_i \ge 1$, then the Pearson χ^2 divergence is also covered. Based on the above result, we consider the following flexible class $\mathcal{S} = \{\hat{p}[\eta, \tau] : \eta \in [\hat{q}_1, \hat{q}_2], 0 \le \tau\}$ of shrinkage prdes, where $\hat{p}[\eta, \tau](\boldsymbol{y}|\boldsymbol{x}) = \prod_i \phi\big(y_i - (\omega_i x_i + \bar{\omega}_i \eta_i);\ (r_i + b\omega_i)\sigma_i^2\,\big)$, with $\omega_i := \omega_i[\tau] = \tau/(\tau + \sigma_i^2)$, $\bar{\omega}_i := \bar{\omega}_i[\tau] = 1 - \omega_i[\tau]$, and $\hat{q}_1, \hat{q}_2$ are $q/2$ and $(1 - q/2)$th quantiles of $X_1, \ldots, X_n$ for any prefixed $q \in (0, 1)$. The class is indexed by the location and scale hyper-parameters, η and τ. For any sensible shrinkage predictors, it is enough to confine the location hyper-parameter η within the $100q\%$ range of the observed data. The scale hyper-parameter τ varies over the non-negative axis. In the following section, we provide a methodology for choosing the hyper-parameters so that the resultant prde has optimal risk properties among all prdes in the class $\mathcal{S}$.

Extension to Non-diagonal Predictive Set-ups. The results and methodology developed here can encompass non-diagonal predictive set-ups where $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)$ and $\boldsymbol{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_f)$ with $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_f$ being known positive definite matrices. For a non-diagonal prior $\boldsymbol{\theta} \sim N(\boldsymbol{\eta}, \Lambda)$, Lemma 2.1 can be extended as follows with α_U being $1 + 2\lambda_{\min}(\boldsymbol{\Sigma}_f \boldsymbol{\Sigma}_p^{-1})$, which is the generalization of the scalar case. The proof of the lemma is presented in section 7.2.

Lemma 2.2. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_p)$, $\boldsymbol{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_f)$, and $\boldsymbol{\theta} \sim N(\boldsymbol{\eta}, \Lambda)$ with known positive definite matrices $\boldsymbol{\Sigma}_p$,$\boldsymbol{\Sigma}_f$, and Λ. Then, the Bayes predictive density estimate with respect to α-divergence loss for $\alpha < \alpha_U$ is*

$$\hat{p}[\boldsymbol{\eta}, \Lambda](\boldsymbol{y} \mid \boldsymbol{x}) = \boldsymbol{\phi}\big(\boldsymbol{y} - (\Omega\, \boldsymbol{x} + \overline{\Omega}\, \boldsymbol{\eta});\, \boldsymbol{\Sigma_f} + b\,(\boldsymbol{\Sigma}_p^{-1} + \Lambda^{-1})^{-1}\big), \tag{2.2}$$

where $\Omega := \Lambda(\Lambda + \boldsymbol{\Sigma}_p)^{-1}$ and $\overline{\Omega} := I - \Omega$.

For tractable shrinkage classes, we need to impose lower dimensional structures on $\boldsymbol{\eta}$ and Λ in (2.2). Perhaps, the most popular choice is $\boldsymbol{\eta} = \eta\mathbf{1}$ and $\Lambda = \tau I$, which extends the class $\mathcal{S}$ based on (2.1) to the non-diagonal set-up. In the following sections, we describe our method and its associated results for the class $\mathcal{S}$. However, note that the methodology can be extended to other shrinkage classes based on (2.2).

Hereon, we describe our method first for the diagonal predictive set-up as it produces comparatively simpler expressions that can be intuitively studied and compared to the point estimation results of Xie et al. [53] and the predictive KL results in Xu and Zhou [56]. The general results for non-diagonal set-ups along with their complete proofs are provided in section 7.

3 Risk Estimation and Hyper-parameter Calibration

Denote by $R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)$ the risk $R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}[\eta, \tau])$ of any arbitrary member $\hat{p}[\eta, \tau]$ in $\mathcal{S}$. The following result proved in section 7.3 shows that this multivariate predictive risk decouples as functions of the corresponding coordinate-wise risks and subsequently can be explicitly written through closed form expressions as functions of $\eta, \tau, \boldsymbol{\theta}$ and α. Letting $\tau \to \infty$, we get the second display in Theorem 2.4 of Ghosh et al. [21].

Theorem 3.1. *For $\alpha \leq \alpha_U$ and $\alpha \neq \pm 1$, the risk of any prde $\hat{p}[\eta, \tau] \in \mathcal{S}$ can be expressed as*

$$\log\big(1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)\big) = \sum_{i=1}^{n} H_i(\theta_i, \eta, \tau; \alpha),\ \ \textit{where, } c_\alpha = (1-\alpha^2)/4,\ \ \textit{and,}$$

$$H_i(\theta_i, \eta, \tau; \alpha) = f\big((\theta_i - \eta)^2, \omega_i[\tau], r_i, \sigma_i, (1-\alpha)/2\big)\ \textit{ with } \omega_i[\tau] = \tau/(\tau + \sigma_i^2),$$

$$f(t, w, r, \sigma, b) = \frac{\bar{b}}{2} \log\left(\frac{r}{r + wb}\right) + \frac{1}{2} \log\left(\frac{r + wb}{r + wb^2 + w^2 b\bar{b}}\right) - \frac{b\bar{b}\bar{w}^2 t}{2\sigma^2(r + wb^2 + w^2 b\bar{b})}$$

and $\bar{w} = 1 - w$, $\bar{b} = 1 - b$.

The risk for the KL and reverse KL losses can be derived from the above expression by noting that for $\alpha = \alpha_0 \in \{-1, 1\}$,

$$R_{n,\alpha_0} = \lim_{\alpha \to \alpha_0} R_{n,\alpha} = \lim_{\alpha \to \alpha_0} \frac{1 - \exp\{\sum_{i=1}^{n} H_i(\theta_i, \eta, \tau; \alpha)\}}{c_\alpha} = 2\alpha_0 \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} H_i(\theta_i, \eta, \tau; \alpha_0),$$

where, the last equality follows from the fact that $H_i = 0$ when $\alpha \in \{-1, 1\}$, and L'Hôpital's rule. Thus,

$$R_{n,-1} = \sum_{i=1}^{n} f_b((\theta_i - \eta)^2, w_i[\tau], r_i, \sigma_i; 1) \text{ and } R_{n,1} = -\sum_{i=1}^{n} f_b((\theta_i - \eta)^2, w_i[\tau], r_i, \sigma_i; 0),$$

where, $f_b(t, w, r, \sigma; b) = \frac{\partial}{\partial b} f(t, w, r, \sigma, b)$. Finally, it yields

$$\begin{aligned} R_{n,-1}(\boldsymbol{\theta}; \eta, \tau) &= \sum_{i=1}^{n} \frac{\log(1 + r_i^{-1}\omega_i)}{2} + \frac{\bar{\omega}_i^2(\theta_i - \eta)^2 - \omega_i\bar{\omega}_i\sigma_i^2}{2\sigma_i^2(r_i + \omega_i)} \ , \\ R_{n,1}(\boldsymbol{\theta}; \eta, \tau) &= \sum_{i=1}^{n} \frac{\bar{\omega}_i^2(\theta_i - \eta)^2 + \omega_i^2\sigma_i^2}{2\,\sigma_i^2 r_i}. \end{aligned}$$

Note that $R_{n,-1}$ matches the KL risk expression in equation (11) of Xu and Zhou [56]. We next estimate the predictive risk $R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)$. Noting that $(X_i - \eta)^2 - \sigma_i^2$ is an unbiased estimator for $(\theta_i - \eta)^2$, an unbiased estimate of $H_i(\theta_i, \eta, \tau; \alpha)$ is given by $\hat{H}_i(\eta, \tau; \alpha)$, where

$$\hat{H}_i(\eta, \tau; \alpha) = f\left((X_i - \eta)^2 - \sigma_i^2 \, , \, \frac{\tau}{\tau + \sigma_i^2} \, , \, r_i \, , \, \sigma_i \, , \, \frac{1 - \alpha}{2} \right) .$$

Consider their average $\hat{\mathcal{H}}_n(\tau, \eta; \alpha) = n^{-1} \sum_{i=1}^{n} \hat{H}_i(\tau, \eta; \alpha)$. For any fixed $\alpha \le \alpha_U$, select the hyper-parameters that minimize the *average risk* estimate, i.e.,

$$(\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}) = \underset{0 \le \tau,\, \eta \in [\hat{q}_1, \hat{q}_2]}{\arg\min} \quad \text{sign}(\alpha^2 - 1)\ \hat{\mathcal{H}}_n(\eta, \tau; \alpha) \tag{3.1}$$

to obtain our proposed prde $\hat{p}[\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}](\boldsymbol{y}|\boldsymbol{x})$. Thus, when $|\alpha| < 1$, we maximize $\hat{\mathcal{H}}_n(\eta, \tau; \alpha)$ and we minimize it when $|\alpha| > 1$. However, note that in both the cases, this corresponds to minimizing risk estimates of the actual risk. For $\alpha = \pm 1$, we directly minimize the unbiased estimates of $R_{n,1}$ and $R_{n,-1}$. Taking the limit of the hyper-parameters $(\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha})$ as $\alpha \to 1_-$ or $\alpha \to -1_+$ also yields similar results.

Figure 1 shows the BH risk of prdes for $n = 5$ and 10 at $\boldsymbol{\theta} = t\mathbf{1}$ where t varies from 0 to ∞. The risks of the best prde in $\mathcal{S}$ (which is characterized later in (4.1)) is plotted in blue. The risk of our proposed method is calculated by Monte-Carlo integration and is plotted in red. In dotted black lines, we have the risk of the best invairant prde $\hat{p}_{\mathsf{U}}$which is the Bayes prde from the uniform prior. We see that compared to $\hat{p}_{\mathsf{U}}$ significant gains in risk can be obtained by estimators in $\mathcal{S}$ when $|t|$ is near the origin and the gains decrease when $|t|$ is large. Also, the risk of the proposed method is reasonably close to the minimum attainable risk in $\mathcal{S}$. Next, we rigorously document these risk properties of prdes tuned by the proposed procedure.

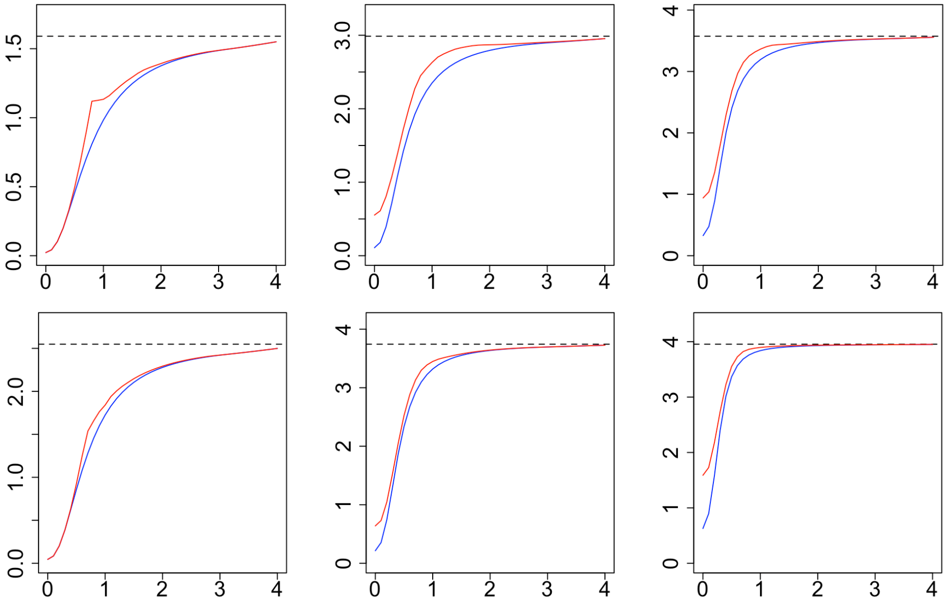

Figure 1: Plot of BH risks of the prdes in a homoskedastic normal model at $\boldsymbol{\theta} = t\mathbf{1}$ as t varies in the abscissa. Here, $\sigma_i = 1$ and $r_i = r$ for all $i = 1, \ldots, n$. From left to right, $r = 1, 0.25, 0.1$ respectively; $n = 5$ in the top and $n = 10$ in the bottom plots. The risk of the best-invariant predictive density (dotted black), the risk of the oracle estimator based on hyper-parameters in (4.1) (in blue) and the risk of our proposed method (in red) are presented here.

4 Theory Results

We first establish a non-asymptotic concentration bound on the deviation of $\hat{\mathcal{H}}_n$ from the true log-risk. Consider the expected absolute deviation:

$$D_{n,\alpha}(\boldsymbol{\theta}, \tau, \eta) = \mathbb{E}\Big[\sup_{\tau \geq 0} \big| n^{-1} \log \left\{1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)\right\} - \hat{\mathcal{H}}_n(\eta, \tau; \alpha) \big| \Big].$$

We establish an upper bound on $D_{n,\alpha}$ that depends on the L_2 norm of the signal strength:

$$g_n(\boldsymbol{\theta}, \eta) := \max \left\{ 1, \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\theta_i - \eta}{\sigma_i} \right)^2 \right\}^{1/2} \right\}.$$

Theorem 4.1. *For any $\alpha \leq \alpha_U$, any fixed $\eta \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^n$, for all $n \geq 1$,*

$$D_{n,\alpha}(\boldsymbol{\theta}, \tau, \eta) \leq \kappa_0\, r_*^{-1} \max\{1, |c_\alpha|\}\, g_n(\boldsymbol{\theta}, \eta)\, n^{-1/2}\ ,$$

where, $\kappa_0 = 12$ is an absolute constant and $r_* = \inf_i r_i$.

When $\alpha \to \pm 1$, both the $D_{n,\alpha}(\boldsymbol{\theta}, \tau, \eta)$ and the $c_\alpha = (1-\alpha^2)/4$ on the RHS tends to 0. Applying L'Hôpital's rule yields the analogous bound $\mathbb{E} \sup_{\tau \geq 0} \left| n^{-1} R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)\} - \hat{\mathcal{H}}_n(\eta, \tau; \alpha)\right| \leq \kappa_0\, r_*^{-1} g_n(\theta, \eta)\, n^{-1/2}$ for $\alpha = \pm 1$. The proof of the theorem is presented in section 7.3.

Our next result which uses Theorem 4.1 shows that our proposed risk estimate approximates the average of the logarithm of the true multivariate risk uniformly well for all prdes in the shrinkage class $\mathcal{S}$. Thus, calibrating the hyper-parameters by minimizing the risk estimates is a sensible choice. To facilitate shorter mathemetical proofs we assume the following asymptotic conditions:

[A1] $\overline{\lim}_{n\to\infty}\, n^{-1} \sum_{i=1}^n \sigma_i^2 < \infty$, $\underline{\lim}_i\, \sigma_i > 0$ and $0 < \underline{\lim}_i\, r_i \leq \overline{\lim}_i\, r_i < \infty$,
[A2] $\overline{\lim}_{n\to\infty}\, n^{-1} \sum_{i=1}^n \theta_i^2 < \infty$.

Though these conditions may be further relaxed as noted after Theorem 3.1 of [53], we do not seek the full generality as the conditions are not restrictive and the proofs presented under these assumptions contain all the essential statistical perspectives.

Theorem 4.2. *Under Assumptions A1-A2, for any $\alpha \leq \alpha_U$ and $a_n = o(n^{1/2})$,*

$$a_n \left(\sup_{\eta \in [\hat{q}_1, \hat{q}_2], \tau \geq 0} \left| n^{-1} \log\left\{1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)\right\} - \hat{\mathcal{H}}_n(\eta, \tau; \alpha)\right| \right) \to 0 \ \textit{in } L_1 \quad \textit{as } n \to \infty\ .$$

The above result (proved in section 7.5) shows that the risk estimates $\hat{\mathcal{H}}_n$ have near-parametric $\sqrt{n}$-rates of convergence barring some poly-log terms. Thus, we expect the risk estimates to be reasonably precise even for moderate dimensions n. This attribute is reflected in the simulation studies in Section 5.

To study the risk properties of our proposed prde $\hat{p}[\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}]$, we next introduce the oracle risk (OR) hyper-parameters as those which minimize the true risk function:

$$(\eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha}) = \operatorname*{arg\,min}_{0 \leq \tau,\ \eta \in [\hat{q}_1, \hat{q}_2]} R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau). \tag{4.1}$$

These oracle choices are not really estimators since they depend on the unknown $\boldsymbol{\theta}$ values. They are not obtainable in practice but provide the theoretical benchmark that one can ever hope to reach. Indeed, no prdes in $\mathcal{S}$ can have smaller risk than the oracle risk prde $\hat{p}[\eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha}]$.

Consider the average logarithmic ratio of the risk deviations

$$\rho_{n,\alpha}(\boldsymbol{\theta}) = -\frac{\gamma_\alpha}{n} \log \left\{ \frac{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha})}{1 - c_\alpha \mathbb{E}\{L_{n,\alpha}(\boldsymbol{\theta}; \hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha})\}} \right\} \text{ where } \gamma_\alpha = \operatorname{sign}(\alpha^2 - 1).$$

By construction, $\rho_{n,\alpha}(\boldsymbol{\theta}) \geq 0$. For any fixed $0 < a_0 < a_1 < \infty$ and $\epsilon > 0$, consider the neighborhood $\Theta[\varepsilon, a_0, a_1] := \{(\eta, \tau) : |\tau - \tau^{\mathsf{or}}_{n,\alpha}| \leq \varepsilon,\, |\eta - \eta^{\mathsf{or}}_{n,\alpha}| \leq \varepsilon,\, \tau \geq 0,\, \eta \in [a_0, a_1]\}$ around oracle hyper-parameters. We impose the following regularity condition on the sensitiveness (non-flatness) on the true risk functions around the oracle hyper-parameters.

[A3] For any $\varepsilon > 0$ and $-\infty < a_0 < a_1 < \infty$,

$$\lim_{n\to\infty} n^{-1/2} \left[\inf_{(\eta,\tau)\notin\Theta[\varepsilon,\, a_0,\, a_1]} \log \left(\frac{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)}{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta^{\mathsf{or}}_{n,\alpha}, \tau^{\mathsf{or}}_{n,\alpha})} \right)^{\gamma_\alpha} \right] = \infty.$$

The following result shows that our estimated hyper-parameters are close to the oracle hyper-parameters and our proposed prde has risk close to the oracle risk. This property is not shared by the EBMLE or EBMOM. The estimating equations for calibrating the hyper-parameters based on the EBMLE and EBMOM (discussed in the following section), differ from our proposed methodology so that the estimated hyper-parameters can be highly different (see cases IV-V in Section 5). In such cases the EBMLE and EBMOM tuned prdes have much higher risk than the oracle unless the risk function is completely flat. As our proposed estimator $\hat{p}[\hat{\eta}_{n,\alpha}, \hat{\tau}_{n,\alpha}]$ is asymptotically nearly as good as the oracle prde, its asymptotic risk is no larger than that of any other prdes in the general class $\mathcal{S}$.

Theorem 4.3. *For any $\alpha \le \alpha_U$, under assumptions A1-A2, the logarithmic ratio $\rho_{n,\alpha}(\boldsymbol{\theta})$ converges asymptotically satisfying $\limsup_{n\to\infty} n^{1/2} \rho_{n,\alpha}(\boldsymbol{\theta}) < \infty$. Additionally, with assumption A3, we have:*

$$(\hat{\eta}_{n,\alpha} - \eta^{\mathsf{or}}_{n,\alpha}, \hat{\tau}_{n,\alpha} - \tau^{\mathsf{or}}_{n,\alpha}) \to 0 \textit{ in probability as } n \to \infty \ .$$

The above result is proved in section 7.6. It shows that as n increases, the average logarithmic ratio of risk deviations from the oracle (ALRORD) converges to 0 for any prde calibrated by the proposed method. As such, as $n \to \infty$ the ALRORD of any prde calibrated by the proposed method is always bound above by $O_p(n^{-1/2})$. Assumption A3 implies that the ALRORD for an arbitrary $\hat{p}[\eta, \tau]$ based on hyper-parameter (η, τ) cannot be bounded below $O_p(n^{-1/2})$ unless the hyper-parameter (η, τ) is within any prefixed $\epsilon-$neighborhood of the oracle hyper-parameter values. This ensures that as $n \to \infty$, the hyper-parameter estimates in (3.1) converge to the oracle values in (4.1).

5 Simulation Experiments

We conduct six simulation experiments to compare the performance of our estimation methodology with competing methods for calibrating estimators in $\mathcal{S}$. We consider the EBMLE tuned prde which uses hyper-parameters

$$(\hat{\eta}_{\mathsf{ML}}, \hat{\tau}_{\mathsf{ML}}) = \operatorname*{arg\,min}_{\eta\in\mathbb{R},\tau\ge 0} \sum_{i=1}^{n} \log(\tau + \sigma_i^2) + \frac{(X_i - \eta)^2}{\tau + \sigma_i^2},$$

the EBMOM tuned prde whose hyper-parameters are solutions to the following equations

$$\hat{\eta}_{\mathsf{MM}} = \frac{\sum_{i=1}^{n} (\sigma_i^2 + \tau)^{-1} X_i}{\sum_{i=1}^{n} (\sigma_i^2 + \tau)^{-1}} \text{ and } \hat{\tau}_{\mathsf{MM}} = \frac{1}{n-1} \left(\sum_{i=1}^{n} (X_i - \hat{\eta}_{\mathsf{MM}})^2 - \frac{n-1}{n} \sum_{i=1}^{n} \sigma_i^2 \right)_+ ,$$

the extended James-Stein [53] based prde in $\mathcal{S}$ with hyper-parameters

$$\hat{\eta}_{\mathsf{JS}} = \frac{\sum_{i=1}^n \sigma_i^{-2} X_i}{\sum_{i=1}^n \sigma_i^{-2}} \text{ and } \hat{\tau}_{\mathsf{JS}} = \left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right)\left(\frac{\sum_{i=1}^n \sigma_i^{-2}(X_i - \hat{\eta}_{\mathsf{JS}})^2}{(n-3)} - 1\right)_+,$$

as well as the completely non-informative (NI) prde which has $\tau \to \infty$ and the oracle estimator of (4.1). Additionally, we also consider the Bayes prde $\hat{p}_{\mathsf{C}}$ from the heavy tailed Cauchy prior.

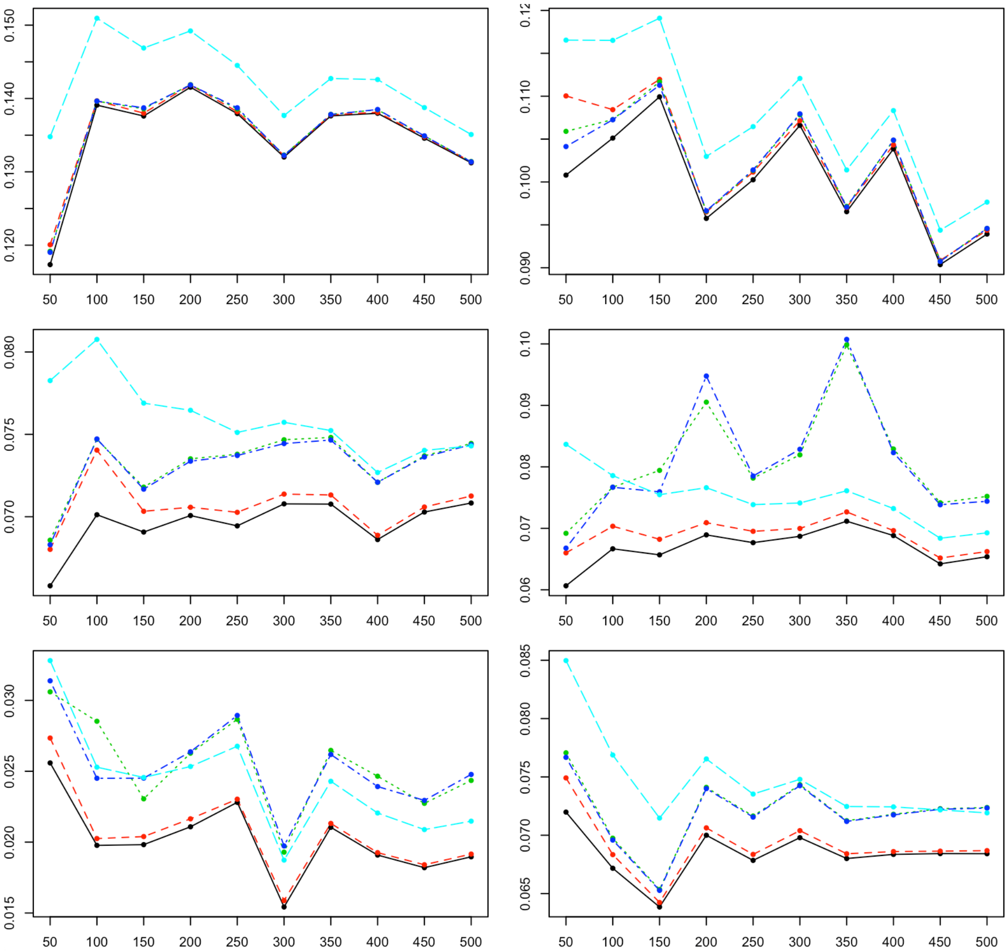

Figure 2: Adjusted BH risk of the prdes based on the oracle estimator of (4.1) (in black), EBMLE (in blue), EBMOM (in green), Cauchy prior based Bayes prde (in cyan) and our proposed method (in red) are plotted as n varies along the x-axis.

The six simulation regimes are inspired by experiments in section 4 of Xie et al. [53]. In Cases I to V, we have Gaussian noise with mean 0 and variances σ_i^2 being i.i.d.

from Uniform$[0.5, 1.5]$ distribution. Thus, the mean noise variance is 1. In Case I, we generate $\boldsymbol{\theta} \sim N(0, 2I_n)$ and the r_i as uniformly distributed between 0.1 and 1. The average signal-to-noise-ratio (snr) is 2 here. Note, that this case is in perfect congruence with the hierarchical normal set-up of Section 2. The remaining cases are different from the normal models and conjugate priors set-ups of Lemmas 2.1 and 2.2. In Case II, θ_i are i.i.d. from a Uniform prior on $[-\sqrt{3}, \sqrt{3}]$ and thus has variability 1. Here, we consider stratification in $\boldsymbol{r}$. The r_i are generated from a mixture of three uniform distributions with significantly different supports. In Case III, we introduce dependence between the means and the future variances by setting $\boldsymbol{\theta} = \boldsymbol{r}$ while $\boldsymbol{r}$ is i.i.d. form Uniform$[0.1, 2]$. Case IV is similar to case 3, except r_i's are no longer bounded as before but are generated from an inverse-Chisquare with 5 degrees of freedom. In Case V, $\{(\theta_i, r_i) : i = 1, \ldots, n\}$ are independent among themselves, r_i takes two possible values with $P(r_i = 10) = 0.7$ and $P(r_i = 1) = 0.3$, and the true mean values are generated conditionally on the r_i values with $[\theta_i | r_i = 10]$ and $[\theta_i | r_i = 1]$ both being normal distributions with mean 0 and standard deviations 0.1 and 1, respectively. In Case VI, we consider $\boldsymbol{\theta}, \boldsymbol{r}$ as in case 3. However, here the noise is no longer Gaussian but is from a uniform distribution with mean 0 and variance 1. The snr is 1.4 in this set-up.

We report the Bhattacharyya-Hellinger predictive risk for the six cases in Table 1 and Figure 2. In each of the six cases, we generate $\{\theta_i, \sigma_i, \nu_i : 1 \leq i \leq n\}$ values once from the corresponding model and then calculate the adjusted BH risk for different $\hat{\eta}$ and $\hat{\tau}$ estimates across 100 replicates. We calculate the BH risk adjusted for dimension by $\text{ABH}_n(\boldsymbol{\theta}; \eta, \tau) = 1 - \{1 - c_0 \, R_{n,0}(\boldsymbol{\theta}; \eta, \tau)\}^{1/n}$. For each case and n, the reported adjusted BH predictive risk in Figure 2 is the average of the ABH value across the 100 replicates.

Table 1: As $n \to \infty$, the limiting BH predictive risk (adjusted for dimensions) of different shrinkage prdes is reported in % of excess risk over that of the oracle estimator in (4.1).

	Proposed	EBMOM	EBMLE	JS	NI	Cauchy
Case I	0.05	0.17	0.19	0.04	24.63	1.63
Case II	0.23	0.71	0.65	0.16	48.30	2.32
Case III	0.21	4.69	4.81	6.70	69.70	2.96
Case IV	0.51	26.82	23.07	26.43	119.47	3.28
Case V	0.55	22.40	23.00	19.00	105.97	9.07
Case VI	0.19	5.70	5.67	5.77	72.54	2.94

In Table 1, $n = 1000$ and so, it reflects the asymptotic risks of these estimators. Figure 2 compares the risk profiles as sample size varies (NI and JS were avoided in the display for their significantly higher error rates reduced clarity in some of the plots). From the table, we see our proposed estimation method is asymptotically close to the oracle in all the cases where as the figure displays that suitable accuracy can be attained for moderate n across all the concerned scenarios. The EBMLE and the EBMOM have risks similar to ours in cases I and II but has significantly worse performance in the other cases when there is dependence between $\boldsymbol{\theta}$ and $\boldsymbol{r}$. The Cauchy prior based Bayes prde $\hat{p}_{\mathsf{C}}$ performs considerably better than the EBMLE and EBMOM in Cases III, IV and V, when n is large (see Table 1), though its risk is still significantly higher than that of the proposed method. In moderate dimensions, the relative risk of $\hat{p}_{\mathsf{C}}$ is substantially higher than the EBMLE, the EBMOM and the proposed methods. As such for $n \leq 200$, the risk curve for $\hat{p}_{\mathsf{C}}$ is higher than the EBMLE or EBMOM in all cases except IV (see figure 2). The JS based prde has erratic asymptotic risk behavior as it often fails to adapt to the heterogeneity in the data and the non-informative prior based prde has poor performance across all regimes. We present two numerical examples in the supplementary materials where suboptimal performance of the JS based prde is witnessed.

6 Discussion and Future Work

We developed a risk estimation based methodology for tuning linearly shrunken prdes in Gaussian models under general α-divergence losses. The proposed risk estimation based method will be particularly useful under heterogeneity. If the set-up is homoskedastic ($\sigma_i = \sigma$ and $\nu_i = \nu$ for all i), then in high dimensions the proposed method will produce hyper-parameter estimates similar to the EBMOM and EBMLE. An interesting topic for future work will be to introspect the roles of prde under α-divergences when covariances are unknown as studied in Kato [27] for the KL loss. Also, in applications it is important to choose an α-divergence loss that is tailored to the specific prediction task at hand [28, 44]. Following Nguyen et al. [41], Gül and Zoubir [23] it will be important to study the roles different α predictive risks in hypothesis testing and classification problems. Another important direction would be understanding the roles of adaptive calibration under α-risk in the presence of latent structures in the mean parameters such as the sparsity restrictions studied in Mukherjee [38], Yano et al. [57] for the KL loss. Finally, extending the risk estimation based methodology developed here to non-normal models will be useful.

7 Proofs

7.1 Details for Non-diagonal Predictive Set-ups

For risk estimation and hyper-parameter calibration in non-diagonal predictive set-ups, we first extend Theorem 3.1. The following result considers any generic $\boldsymbol{\eta} \in \mathbb{R}^n$ and $n \times n$ positive definite (pd) matrix Λ and provides the expression for α-risk of Bayes prdes given by (2.2).

Lemma 7.1. *For $\alpha \neq \pm 1$ and $\alpha \leq \alpha_U$,*

$$R_{n,\alpha}(\boldsymbol{\theta}, \hat{p}[\boldsymbol{\eta}, \Lambda]) = (b\bar{b})^{-1}\big\{1 - F(b)\exp\big[-b\bar{b}\boldsymbol{t}^\top\overline{\Omega}^\top(\boldsymbol{\Sigma}_f + b^2\Omega\boldsymbol{\Sigma}_p + b\bar{b}\Omega^\top\boldsymbol{\Sigma}_p\Omega)^{-1}\overline{\Omega}\boldsymbol{t}/2\big]\big\}, \tag{7.1}$$

where $\boldsymbol{t} := \boldsymbol{\theta} - \boldsymbol{\eta}$, $\Omega := \Lambda(\Lambda + \boldsymbol{\Sigma}_p)^{-1}$, $b = (1-\alpha)/2$ and,

$$F(b) := \left(\frac{\det(\boldsymbol{\Sigma}_f)}{\det(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)}\right)^{\bar{b}/2}\left(\frac{\det(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)}{\det(\boldsymbol{\Sigma}_f + b^2\Omega\boldsymbol{\Sigma}_p + b\bar{b}\Omega^\top\boldsymbol{\Sigma}_p\Omega)}\right)^{1/2}.$$

Also, $2R_{n,1}(\boldsymbol{\theta}, \hat{p}[\boldsymbol{\eta}, \Lambda]) = \mathrm{tr}(\boldsymbol{\Sigma}_f^{-1}\Omega^\top \boldsymbol{\Sigma}_p \Omega) + \boldsymbol{t}^\top \overline{\Omega}^\top \boldsymbol{\Sigma}_f^{-1}\overline{\Omega}\boldsymbol{t}$ *and* $2R_{n,-1}(\boldsymbol{\theta}, \hat{p}[\boldsymbol{\eta}, \Lambda])$ *equals*

$$\log\det(I + \boldsymbol{\Sigma}_f^{-1}\Omega\boldsymbol{\Sigma}_p) + \boldsymbol{t}^\top \overline{\Omega}^\top (\boldsymbol{\Sigma}_f + \Omega\boldsymbol{\Sigma}_p)^{-1}\overline{\Omega}\boldsymbol{t} - \mathrm{tr}\{(\boldsymbol{\Sigma}_f + \Omega\boldsymbol{\Sigma}_p)^{-1}(\Omega\boldsymbol{\Sigma}_p - \Omega^\top\boldsymbol{\Sigma}_p\Omega)\}.$$

The above expression (7.1) involves the hyper-parameters in Λ via Ω and $\bar{\Omega}$; it is just the matrix version of risk expression in Theorem 3.1 and was obtained in the above closed form by applying the Sherman-Morrison-Woodbury formula to the multi-variate risk function expressions. The proof is provided later in section 7.3.

Next, we consider hyper-parameter calibration for the prdes $\hat{p}[\boldsymbol{\eta}, \Lambda]$ which are given by (2.2) and whose risks are given by (7.1). For that purpose, assume $\boldsymbol{\eta} = \eta\mathbf{1}$ and $\Lambda = \tau I$. Also, assume that $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_f$ has same eigen-vectors, i.e., $\boldsymbol{\Sigma}_p = \boldsymbol{U}\mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}\boldsymbol{U}^\top$ and $\boldsymbol{\Sigma}_f = \boldsymbol{U}\mathrm{diag}\{\nu_1^2, \ldots, \nu_n^2\}\boldsymbol{U}^\top$ with a known orthogonal matrix $\boldsymbol{U} = [U_1, \ldots, U_n]$.

Define $u_i = \mathbf{1}^T U_i$ for $i = 1, \ldots, n$. Consider generalization of the following definitions from section 3 of the main paper:

$$\begin{aligned} &H_i(\theta_i, \eta, \tau; \alpha) = f\big((\theta_i - \eta\, u_i)^2, \omega_i[\tau], r_i, \sigma_i, (1-\alpha)/2\big) \text{ and} \\ &\hat{H}_i(\eta, \tau; \alpha) = f\big((U_i^T\boldsymbol{X} - \eta\, u_i)^2 - \sigma_i^2,\ \tau/(\tau + \sigma_i^2), r_i, \sigma_i, (1-\alpha)/2\big), \end{aligned}$$

where, f, $\omega[\tau]$, r_i are defined as before. Note, that substituting $\boldsymbol{U} = I$ above, we get back the definitions in Section 3.

Noting that α-divergence risk is invariant with respect to any one-to-one transformations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, consider the transformed prediction problem with $\widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{X} \sim N(\widetilde{\boldsymbol{\theta}}, \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\})$, $\widetilde{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{Y} \sim N(\widetilde{\boldsymbol{\theta}}, \mathrm{diag}\{\nu_1^2, \ldots, \nu_n^2\})$ and $\widetilde{\boldsymbol{\theta}} = \boldsymbol{U}\boldsymbol{\theta}$. The prior structure on $\widetilde{\boldsymbol{\theta}} = \{\widetilde{\theta}_i : 1 \leq i \leq n\}$ is $\widetilde{\theta}_i \sim N(\eta\, u_i, \tau)$ independently for $i = 1, \ldots, n$. With this transformed set-up, from lemmas 2.2 and 7.1 it follows that theorem 4.1 can be easily extended to

$$\log(1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta}; \eta, \tau)) = \sum_{i=1}^n H_i(\widetilde{\theta}_i, \eta, \tau; \alpha) = \mathbb{E}\bigg\{\sum_{i=1}^n \hat{H}_i(\eta, \tau; \alpha)\bigg\}.$$

Thus, the optimization in (3.1) is also our prescribed method for the non-diagonal predictive set-up. With the additional assumption $\limsup_i \sigma_i < \infty$, the optimality results of section 4 also extend over with the general definition

$$g_n(\boldsymbol{\theta}, \eta) = \max\left\{1, \left\{\frac{1}{n}\sum_{i=1}^n \left(\frac{U_i^T\boldsymbol{\theta} - \eta \cdot U_i^T\mathbf{1}}{\sigma_i}\right)^2\right\}^{1/2}\right\}.$$

Next, we provide the proofs of the results stated in the main paper. We provide the proofs for the general non-diagonal set-up from which the results for the diagonal case directly follows. Note that, we do need the assumption that $\boldsymbol{\Sigma}_f$ and $\boldsymbol{\Sigma}_p$ have same eigen vectors for Lemma 2.2 and 7.1 but for the generalized versions of the optimality results of Section 4.

7.2 Proofs of Lemmas 2.1 and 2.2

We begin with two additional notations: let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ respectively be the maximum and the minimum eigenvalue of a matrix M. For positive definite (pd) M, let $\boldsymbol{\phi}(\cdot; M)$ denote multivariate normal probability density function with covariance matrix M. We prove lemma 2.2 from which lemma 2.1 easily follows.

Consider the cases when $\alpha \neq \pm 1$ first. This is equivalent to $b \neq 0, 1$. Note, that $\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p$ is pd. We start by expanding $\boldsymbol{\phi}(\boldsymbol{y} - (\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta}); \boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)$. Let

$$T_1 := \boldsymbol{y}^\top(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)^{-1}\boldsymbol{y} \text{ and } T_2 := \boldsymbol{y}^\top(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)^{-1}(\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta}).$$

We expand $\{\boldsymbol{y} - (\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta})\}^\top(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)^{-1}\{\boldsymbol{y} - (\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta})\}$ as

$$T_1 - 2T_2 + (\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta})^\top(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)^{-1}(\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta}) \ . \tag{7.2}$$

Next, consider further reduction of T_1 and T_2. By Sherman–Morrison–Woodbury formula we know

$$(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)^{-1} = \boldsymbol{\Sigma}_f^{-1} - \boldsymbol{\Sigma}_f^{-1}(\boldsymbol{\Sigma}_f^{-1} + b^{-1}\boldsymbol{\Sigma}_p^{-1} + b^{-1}\Lambda^{-1})^{-1}\boldsymbol{\Sigma}_f^{-1} \ ,$$

using which we get

$$T_1 = \boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}(\boldsymbol{\Sigma}_f^{-1} + b^{-1}\boldsymbol{\Sigma}_p^{-1} + b^{-1}\Lambda^{-1})^{-1}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}. \tag{7.3}$$

A simple algebra shows $(\boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)^{-1}\Omega = \boldsymbol{\Sigma}_f\{b\boldsymbol{\Sigma}_f^{-1} + \boldsymbol{\Sigma}_p^{-1} + \Lambda^{-1}\}^{-1}\boldsymbol{\Sigma}_p^{-1}$ and thus we get

$$T_2 = \boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\{\boldsymbol{\Sigma}_f^{-1} + b^{-1}\boldsymbol{\Sigma}_p^{-1} + b^{-1}\Lambda^{-1}\}(b^{-1}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x} + b^{-1}\Lambda^{-1}\boldsymbol{\eta}) \tag{7.4}$$

Combined with (7.2), equations (7.3) and (7.4) yield $\boldsymbol{\phi}(\boldsymbol{y} - (\Omega\boldsymbol{x} + \overline{\Omega}\boldsymbol{\eta}); \boldsymbol{\Sigma}_f + b\Omega\boldsymbol{\Sigma}_p)$

$$\begin{aligned}\propto \exp[&-\boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}/2 + \boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}(\boldsymbol{\Sigma}_f^{-1} + b^{-1}\boldsymbol{\Sigma}_p^{-1} + b^{-1}\Lambda^{-1})^{-1}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}/2 \\ &+ \boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\{\boldsymbol{\Sigma}_f^{-1} + b^{-1}\boldsymbol{\Sigma}_p^{-1} + b^{-1}\Lambda^{-1}\}(b^{-1}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x} + b^{-1}\Lambda^{-1}\boldsymbol{\eta})],\end{aligned} \tag{7.5}$$

where $\propto$ denotes equality up to a multiplicative constant independent of $\boldsymbol{y}$.

We next expand $\hat{p}[\boldsymbol{\eta}, \Lambda]$. To this end, we employ the following representation of the Bayes prde based on prior $\pi_n(\boldsymbol{\theta})$: If $\{\int \boldsymbol{\phi}^b(\boldsymbol{y} - \boldsymbol{\theta}; \boldsymbol{\Sigma}_f)\boldsymbol{\phi}(\boldsymbol{x} - \boldsymbol{\theta}; \boldsymbol{\Sigma}_p)\pi_n(\boldsymbol{\theta})d\boldsymbol{\theta}\}^{1/b}$ is integrable with respect to $\boldsymbol{y}$ for all $\boldsymbol{x}$, then

$$\hat{p}[\boldsymbol{\eta}, \Lambda](\boldsymbol{y} \mid \boldsymbol{x}) \propto \left\{\int \boldsymbol{\phi}^b(\boldsymbol{y} - \boldsymbol{\theta}; \boldsymbol{\Sigma}_f)\boldsymbol{\phi}(\boldsymbol{x} - \boldsymbol{\theta}; \boldsymbol{\Sigma}_p)\pi_n(\boldsymbol{\theta})d\boldsymbol{\theta}\right\}^{1/b} . \tag{7.6}$$

For the proof of this representation, see Theorem 1 of [9] and remarks therein. Consider the right hand side in (7.6). Observe that for any positive definite matrix A and any n-dimensional vector $\boldsymbol{b}$, we have

$$\int \exp\{-\boldsymbol{\theta}^\top A^{-1}\boldsymbol{\theta}/2 + \boldsymbol{\theta}^\top\mathbf{b}\}d\boldsymbol{\theta} = (2\pi)^{n/2}(\det(A))^{1/2}\exp\{\boldsymbol{b}^\top A\boldsymbol{b}/2\}. \tag{7.7}$$

Using the above, we arrive at

$$
\begin{aligned}
&\left\{\int \boldsymbol{\phi}^b(\boldsymbol{y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)\boldsymbol{\phi}(\boldsymbol{x}-\boldsymbol{\theta};\boldsymbol{\Sigma}_p)\boldsymbol{\phi}(\boldsymbol{\theta}-\boldsymbol{\eta};\Lambda)d\boldsymbol{\theta}\right\}^{1/b}\\
&\propto \mathrm{e}^{-\boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}/2}\left[\int \exp\left\{-\frac{\boldsymbol{\theta}^\top(b\boldsymbol{\Sigma}_f^{-1}+\boldsymbol{\Sigma}_p^{-1}+\Lambda^{-1})\boldsymbol{\theta}}{2}+\boldsymbol{\theta}^\top(b\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}+\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x}+\Lambda^{-1})\right\}d\boldsymbol{\theta}\right]^{1/b}\\
&\propto \exp\{-\boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}/2\\
&\qquad +(b\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}+\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x}+\Lambda^{-1})^\top(b\boldsymbol{\Sigma}_f^{-1}+\boldsymbol{\Sigma}_p^{-1}+\Lambda^{-1})^{-1}(b\boldsymbol{\Sigma}_f^{-1}\boldsymbol{y}+\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x}+\Lambda^{-1})/2\}\\
&\propto \exp[-\boldsymbol{y}^\top\{\boldsymbol{\Sigma}_f^{-1}-\boldsymbol{\Sigma}_f^{-1}(\boldsymbol{\Sigma}_f^{-1}+b^{-1}\boldsymbol{\Sigma}_p^{-1}+b^{-1}\Lambda^{-1})^{-1}\boldsymbol{\Sigma}_f^{-1}\}\boldsymbol{y}/2\\
&\qquad +\boldsymbol{y}^\top\boldsymbol{\Sigma}_f^{-1}\{\boldsymbol{\Sigma}_f^{-1}+b^{-1}\boldsymbol{\Sigma}_p^{-1}+b^{-1}\Lambda^{-1}\}(b^{-1}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{x}+b^{-1}\Lambda^{-1}\boldsymbol{\eta})].
\end{aligned}
$$

Together with (7.5) and (7.6), this shows that if $\{\int \boldsymbol{\phi}^b(\boldsymbol{y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)\boldsymbol{\phi}(\boldsymbol{x}-\boldsymbol{\theta};\boldsymbol{\Sigma}_p)\boldsymbol{\phi}(\boldsymbol{\theta}-\boldsymbol{\eta};\Lambda)d\boldsymbol{\theta}\}^{1/b}$ is integrable with respect to $\boldsymbol{y}$ for all $\boldsymbol{x}$, we have

$$
\hat{p}[\boldsymbol{\eta},\Lambda](\boldsymbol{y}\mid\boldsymbol{x}) \propto \boldsymbol{\phi}(\boldsymbol{y}-(\Omega\boldsymbol{x}+\overline{\Omega}\boldsymbol{\eta});\boldsymbol{\Sigma}_f+b\Omega\boldsymbol{\Sigma}_p). \tag{7.8}
$$

Lastly, we will check the condition that $\{\int \boldsymbol{\phi}^b(\boldsymbol{y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)\boldsymbol{\phi}(\boldsymbol{x}-\boldsymbol{\theta};\boldsymbol{\Sigma}_p)\boldsymbol{\phi}(\boldsymbol{\theta}-\boldsymbol{\eta};\Lambda)d\boldsymbol{\theta}\}^{1/b}$ is integrable with respect to $\boldsymbol{y}$ for all $\boldsymbol{x}$. Equation (7.8) also implies that if $\boldsymbol{\Sigma}_f+b\Omega\boldsymbol{\Sigma}_p$ is positive definite, then $\{\int \boldsymbol{\phi}^b(\boldsymbol{y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)\boldsymbol{\phi}(\boldsymbol{x}-\boldsymbol{\theta};\boldsymbol{\Sigma}_p)\boldsymbol{\phi}(\boldsymbol{\theta}-\boldsymbol{\eta};\Lambda)d\boldsymbol{\theta}\}^{1/b}$ is integrable with respect to $\boldsymbol{y}$ for all $\boldsymbol{x}$. Note that under the assumption that $\alpha < 1+2\lambda_{\min}(\boldsymbol{\Sigma}_f\boldsymbol{\Sigma}_p^{-1})$, matrix $\boldsymbol{\Sigma}_f+b\Omega\boldsymbol{\Sigma}_p$ is positive definite. This completes the proof. When $\alpha=\pm 1$, the proof is much easier and follows by repeating part of the above deductions.

7.3 Proofs for Theorem 3.1 and Lemma 7.1

We prove Lemma 7.1 from which Theorem 3.1 directly follows. In addition to the notations defined at the beginning of subsection 7.2, define

$$
H_\alpha(\boldsymbol{\theta},\hat{p}) := \int \boldsymbol{\phi}(\boldsymbol{x}-\boldsymbol{\theta};\boldsymbol{\Sigma}_p)\int \boldsymbol{\phi}(\boldsymbol{y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)\left\{\frac{\hat{p}(\boldsymbol{y}\mid\boldsymbol{x})}{\boldsymbol{\phi}(\boldsymbol{y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)}\right\}^{\bar{b}}
$$

for any prde $\hat{p}$ and for $\alpha \neq \pm 1$. We start with the case with $\alpha \neq \pm 1$. For a square matrix $\boldsymbol{A}$, an n-dimensional vector $\boldsymbol{b}$, and a positive definite matrix $\boldsymbol{M}$, define the prde $\hat{p}[\boldsymbol{A},\boldsymbol{b},\boldsymbol{M}] := \boldsymbol{\phi}(\boldsymbol{y}-\{\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}\};\boldsymbol{M})$. It suffices to show that, if both $bI+\bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2}$ and $I+\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}^\top\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_p^{1/2}$ with $\boldsymbol{S} := \{b\boldsymbol{M}+\bar{b}\boldsymbol{\Sigma}_f\}/(b\bar{b})$ are positive definite, then

$$
\begin{aligned}
\log H_\alpha(\boldsymbol{\theta},\hat{p}[\boldsymbol{A},\boldsymbol{b},\boldsymbol{M}]) &= \frac{\bar{b}}{2}\log\left(\frac{\det(\boldsymbol{\Sigma}_f)}{\det(\boldsymbol{M})}\right)-\frac{1}{2}\log\det(bI+\bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2})\\
&-\frac{1}{2}\log\det(I+\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}^\top\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_p^{1/2})-\frac{1}{2}(\theta-\boldsymbol{A}\theta-\boldsymbol{b})^\top(\boldsymbol{S}+\boldsymbol{A}^\top\boldsymbol{\Sigma}_p\boldsymbol{A})^{-1}(\theta-\boldsymbol{A}\theta-\boldsymbol{b}).
\end{aligned}
\tag{7.9}
$$

Together with the identity $R_{n,\alpha}(\boldsymbol{\theta},\hat{p}[\boldsymbol{\eta},\Lambda]) = \{1-H_\alpha(\boldsymbol{\theta},\hat{p}[\boldsymbol{\eta},\Lambda])\}/(b\bar{b})$, this gives the desired expression of $R_{n,\alpha}$ by setting $\boldsymbol{A}=\Omega$, $\boldsymbol{b}=\bar{\Omega}\eta$, and $\boldsymbol{M}=\boldsymbol{\Sigma}_f+b\Omega\boldsymbol{\Sigma}_p$, by noting that $b\boldsymbol{M} + \bar{b}\boldsymbol{\Sigma}_f = \boldsymbol{\Sigma}_f + b^2\Omega\boldsymbol{\Sigma}_p$ and by noting that the Sylvester determinant theorem yields

$$\begin{aligned}
&\det(bI + \bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2}) \times \det(I + \boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}_p^{1/2}) \\
&= \det(bI + \bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2}) \times \frac{\det(\boldsymbol{S} + \boldsymbol{A}^\top\boldsymbol{\Sigma}_p A)}{\det(\boldsymbol{S})} \\
&= \frac{\det(b\boldsymbol{M} + \bar{b}\boldsymbol{\Sigma}_f)}{\det(\boldsymbol{M})} \times \frac{\det(\boldsymbol{S} + \boldsymbol{A}^\top\boldsymbol{\Sigma}_p\boldsymbol{A})}{\det(\boldsymbol{S})}.
\end{aligned}$$

Hereafter we denote by $\mathrm{E}_{\boldsymbol{X}|\boldsymbol{\theta}}$ and $\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}$ the expectations with respect to $\boldsymbol{\phi}(\boldsymbol{X}-\boldsymbol{\theta};\boldsymbol{\Sigma}_f)$ and $\boldsymbol{\phi}(\boldsymbol{Y}-\boldsymbol{\theta};\boldsymbol{\Sigma}_p)$, respectively. Observing that $\bar{b}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^\top\boldsymbol{M}^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) - (\boldsymbol{y}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}_f^{-1}(\boldsymbol{y}-\boldsymbol{\theta})$ equals $\bar{b}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\boldsymbol{M}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta}) + \bar{b}(\boldsymbol{y}-\boldsymbol{\theta})^\top(\boldsymbol{M}^{-1} - \boldsymbol{\Sigma}_f^{-1})(\boldsymbol{y}-\boldsymbol{\theta}) - 2\bar{b}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\boldsymbol{M}^{-1}(\boldsymbol{y}-\boldsymbol{\theta})$, it follows that $H_\alpha(\boldsymbol{\theta},\hat{p}[\boldsymbol{A},\boldsymbol{b},\boldsymbol{M}])$ equals

$$\left(\frac{\det(\boldsymbol{\Sigma}_f)}{\det(\boldsymbol{M})}\right)^{\bar{b}/2} \mathrm{E}_{\boldsymbol{X}|\boldsymbol{\theta}}\left[\exp\left\{-\frac{\bar{b}}{2}(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}-\boldsymbol{\theta})^\top\boldsymbol{M}^{-1}(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}-\boldsymbol{\theta})\right\} I(\boldsymbol{X})\right], \text{ where,} \tag{7.10}$$

$I(\boldsymbol{x})$ is defined as

$$\mathrm{E}_{\boldsymbol{Y}|\boldsymbol{\theta}}\left[\exp\left\{-\frac{\bar{b}}{2}(\boldsymbol{Y}-\boldsymbol{\theta})^\top(\boldsymbol{M}^{-1}-\boldsymbol{\Sigma}_f^{-1})(\boldsymbol{Y}-\boldsymbol{\theta}) + \bar{b}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\boldsymbol{M}^{-1}(\boldsymbol{Y}-\boldsymbol{\theta})\right\}\right].$$

Formula (7.7) yields $I(\boldsymbol{x})$ equals

$$\begin{aligned}
&\int \frac{\exp\{-\boldsymbol{z}^\top[bI + \bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2}]\boldsymbol{z}/2\}}{(2\pi)^{n/2}} \exp\{\bar{b}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{z}\}d\boldsymbol{z} \\
&= \frac{\exp\{\bar{b}^2(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2}(bI + \bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2})^{-1}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})/2\}}{\{\det(bI + \bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2})\}^{1/2}}.
\end{aligned}$$

Putting this into (7.10), we get

$$H_\alpha(\boldsymbol{\theta},\hat{p}[\boldsymbol{A},\boldsymbol{b},\boldsymbol{M}]) = \left(\frac{\det(\boldsymbol{\Sigma}_f)}{\det(\boldsymbol{M})}\right)^{\bar{b}/2} \frac{\mathrm{E}_{\boldsymbol{X}|\boldsymbol{\theta}}[\exp\{-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})\}]}{\{\det(bI + \bar{b}\boldsymbol{\Sigma}_f^{1/2}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_f^{1/2})\}^{1/2}}$$

where, $\widetilde{\boldsymbol{S}}^{-1} := \bar{b}\boldsymbol{M}^{-1} - \bar{b}^2\boldsymbol{M}^{-1}\{b\boldsymbol{\Sigma}_f^{-1} + \bar{b}\boldsymbol{M}^{-1}\}^{-1}\boldsymbol{M}^{-1}$.

Lastly, we evaluate $\mathrm{E}_{\boldsymbol{X}|\boldsymbol{\theta}}[\exp\{-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})\}$. Observe that $(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})$ can be decomposed as

$$\begin{aligned}
(\boldsymbol{x}-\boldsymbol{\theta})^\top\boldsymbol{A}^\top\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\theta}) - 2(\boldsymbol{x}-\boldsymbol{\theta})^\top\boldsymbol{A}^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b}) \\
+(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b}).
\end{aligned}$$

Together with (7.10), this implies that $\mathrm{E}_{\boldsymbol{X}|\boldsymbol{\theta}}[\exp\{-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})\}$ is given by the product of: $I_1 := \exp\{-(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})/2\}$, $I_2 := \{\det(I+\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}^\top\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_p^{1/2})\}^{-1/2}$, and $I_3 := \exp\{(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})^\top\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}_p^{1/2}(I+\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}^\top\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_p^{1/2})^{-1}\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})/2\}$.

Next, applying the following Sherman–Morrison–Woodbury decomposition

$$(\widetilde{\boldsymbol{S}}+\boldsymbol{A}^\top\boldsymbol{\Sigma}_p\boldsymbol{A})^{-1} = \widetilde{\boldsymbol{S}}^{-1} - \widetilde{\boldsymbol{S}}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}_p^{1/2}(I+\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}\widetilde{\boldsymbol{S}}^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}_p^{1/2})^{-1}\boldsymbol{\Sigma}_p^{1/2}\boldsymbol{A}\widetilde{\boldsymbol{S}}^{-1} \ ,$$

we reduce $I_1 \cdot I_3$ to simplify the above product and eventually arrive at the conclusion that $\mathrm{E}_{\boldsymbol{X}|\boldsymbol{\theta}}[\exp\{-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})^\top\widetilde{\boldsymbol{S}}^{-1}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}-\boldsymbol{\theta})\}$ equals

$$I_2 \ \cdot \exp\{-(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})^\top(\widetilde{\boldsymbol{S}}+\boldsymbol{A}^\top\boldsymbol{\Sigma}_p\boldsymbol{A})^{-1}(\boldsymbol{\theta}-\boldsymbol{A}\boldsymbol{\theta}-\boldsymbol{b})/2\} \ . \tag{7.11}$$

Further, the Sherman–Morrison–Woodbury formula also yields $\widetilde{\boldsymbol{S}} = \boldsymbol{S}$. Thus, we obtain the desired identity (7.9) for $\alpha \neq \pm 1$.

Finally, we consider the cases with $\alpha = \pm 1$. These cases can be obtained by taking the limit of $R_{n,\alpha}$. Noting that

$$\frac{F'(b)}{F(b)} = \Bigg[\frac{\log\det(I+\bar{b}\boldsymbol{\Sigma}_f^{-1}\Omega\boldsymbol{\Sigma}_p)}{2} + \frac{\bar{b}\mathrm{tr}(\boldsymbol{\Sigma}_f^{-1}\Omega\boldsymbol{\Sigma}_p)}{2} + \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_f+b\Omega\boldsymbol{\Sigma}_p)^{-1}\Omega\boldsymbol{\Sigma}_p\}}{2} \\ - \frac{\mathrm{tr}\{(\boldsymbol{\Sigma}_f+b^2\Omega\boldsymbol{\Sigma}_p+b\bar{b}\Omega^\top\boldsymbol{\Sigma}_p\Omega)^{-1}(2b\Omega\boldsymbol{\Sigma}_p-2b\Omega^\top\boldsymbol{\Sigma}_p\Omega+\Omega^\top\boldsymbol{\Sigma}_p\Omega)\}}{2}\Bigg],$$

L'Hôpital's rule gives the expressions for $\alpha = \pm 1$, which completes the proof.

7.4 Proof for Theorem 4.1

By the translation invariance of an α-divergence, we set $\sigma_i := 1$. Let $\varepsilon_i := (X_i - \theta_i)$ for $i = 1, \ldots, n$. Let $\omega(\tau) := \tau/(\tau+1)$ and $\bar{\omega}(\tau) := 1 - \omega(\tau)$. Let $\alpha \neq \pm 1$ first. Start with the following Layer cake representation of the target quantity:

$$\begin{aligned}&\mathbb{E}\Bigg[\sup_{\tau\geq 0}\left|n^{-1}\log\left\{1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right|\Bigg] \\ &\quad = \int_0^\infty \Pr\left[\sup_{\tau\geq 0}\left|n^{-1}\log\left\{1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right| > x\right]dx.\end{aligned}$$

Let $m_i(\tau) := \bar{\omega}^2(\tau)/(r_i + b^2\omega(\tau) + b\bar{b}\omega^2(\tau))$ for all $i = 1, \ldots, n$ and $\tau \in [0,\infty)$. We upper bound the target quantity $|n^{-1}\log\left\{1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)|$ by the sum of two tractable quantities

$$\frac{|b\bar{b}|}{n}\left|\sum_{i=1}^n\{(X_i-\theta_i)^2-1\}m_i(\tau)\right| + \frac{|b\bar{b}|}{n}\left|\sum_{i=1}^n(\theta_i-\eta)(X_i-\theta_i)m_i(\tau)\right|.$$

Thus, we have

$$\mathbb{E}\left[\sup_{\tau\geq 0}\left|n^{-1}\log\left\{1-c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\}-\hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right|\right]\leq\int_0^\infty\left\{P_1(x)+P_1(x)\right\}dx, \tag{7.12}$$

$$\text{where, } P_1(x)=\Pr\left(\sup_{\tau\geq 0}\left|\sum_{i=1}^n(\varepsilon_i^2-1)m_i(\tau)\right|>\frac{nx}{2|b\bar{b}|}\right)\text{ and,}$$

$$P_2(x)=\Pr\left(\sup_{\tau\geq 0}\left|\sum_{i=1}^n(\theta_i-\eta)\varepsilon_i m_i(\tau)\right|>\frac{nx}{2|b\bar{b}|}\right).$$

Consider bounding $P_1(x)$. Without loss of generality, we assume $r_1\geq r_2\geq r_3\geq\cdots\geq r_n$. Noting that $m_i(\tau)=(\tau^2(r_i+b)+\tau(2r_i+b^2)+r_i)^{-1}\leq r_i^{-1}$, it follows that for any $\tau\in[0,\infty)$, we have $0\leq m_1(\tau)\leq m_2(\tau)\leq\cdots\leq m_n(\tau)\leq\underline{r}^{-1}$ with $\underline{r}:=\inf r_i$. Now, defining a standardized version $\tilde{m}_i(\tau)$ of $m_i(\tau)$ as $\tilde{m}_i(\tau):=\underline{r}m_i(\tau)$, we arrive at

$$\sup_{\tau\geq 0}\left|\sum_{i=1}^n(\varepsilon_i^2-1)m_i(\tau)\right|=\frac{1}{\underline{r}}\sup_{\tau\geq 0}\sum_{i=1}^n|\varepsilon_i^2-1|\tilde{m}_i(\tau) \tag{7.13}$$

which is bounded above by the magnitude of the following weighted sum

$$\underline{r}^{-1}\sup\left\{\sum_{i=1}^n|\varepsilon_i^2-1|a_i:\boldsymbol{a}\in\mathbb{R}^n\text{ and }0\leq a_1\leq a_2\leq\cdots\leq a_n\leq 1\right\}.$$

The function $(a_1,\ldots,a_n)\mapsto\sum_{i=1}^n a_i|\varepsilon_i^2-1|$ is convex, and the set of extreme points of $\{\boldsymbol{a}\in\mathbb{R}^n:0\leq a_1\leq a_2\leq\cdots\leq a_n\leq 1\}$ is exactly $\{\boldsymbol{a}\in\mathbb{R}^n:a_j=\cdots=a_n=1\text{ for some }1\leq j\leq n\}$. So, the maximum of the above quantity is attained by $a_{j^*}=\cdots=a_n=1$ for some j^*. Thus, the expression in (7.13) equals the following simpler quantity: $\underline{r}^{-1}\max_{1\leq j\leq n}\sum_{i=j}^n|\varepsilon_i^2-1|$.

With this simplified expression, we now apply the Kolmogorov inequality, to bound

$$P_1(x)\leq\left(\frac{2|b\bar{b}|}{nx\underline{r}}\right)^2\left(\sum_{j=1}^n\mathbb{E}(\varepsilon_i^2-1)^2\right)\leq\frac{1}{n}\left(\frac{2|b\bar{b}|}{x\,\underline{r}}\right)^2\left\{\mathbb{E}\varepsilon^4-1\right\}=\frac{1}{nx^2}C_1\ ,$$

where ε has the standard normal distribution and $C_1:=8b^2\bar{b}^2\underline{r}^{-2}$. Next, using the inequality between arithmetic and geometric means, it follows

$$\int_0^\infty P_1(x)dx\leq\inf_{x>0}\{x+C_1n^{-1}x^{-1}\}=2C_1^{1/2}n^{-1/2}. \tag{7.14}$$

Next consider bounding $P_2(x)$. Similarly as before, it follows that

$$\sup_{\tau\geq 0}\left|\sum_{i=1}^n\varepsilon_i(\theta_i-\eta)m_i(\tau)\right|\leq\frac{1}{\underline{r}}\sup_{1\leq j\leq n}\left|\sum_{i=j}^n\varepsilon_i(\theta_i-\eta)\right|.$$

As in bounding $P_1(x)$, here also we apply the Markov and the Rosenthal inequalities. It yields

$$
\begin{aligned}
P_2(x) &\le \left(\frac{2|b\bar{b}|/\underline{r}}{nx}\right)^4 \mathbb{E} \sup_{1\le j\le n} \left|\sum_{i=j}^{n} \varepsilon_i(\theta_i-\eta)\right|^4 \\
&\le \left(\frac{2|b\bar{b}|/\underline{r}}{nx}\right)^4 C_4 \left[\mathbb{E}|\varepsilon|^4 \frac{1}{n}\sum_{i=1}^{n}|\theta_i-\eta|^4 + \left\{\frac{1}{n}\sum_{i=1}^{n}(\theta_i-\eta)^2\right\}^2\right] \\
&\le \frac{(2|b\bar{b}|/\underline{r})^4 C_4}{n^2x^4}(\mathbb{E}|\varepsilon_1|^4+1)\left\{\frac{1}{n}\sum_{i=1}^{n}(\theta_i-\eta)^2\right\}^2. \qquad (7.15)
\end{aligned}
$$

Here, C_4 is the constant in the Rosenthal inequality for the fourth moment [25]; $C_4 = \mathbb{E}(W-1)^4$ where W has Poisson distribution with mean 1. Thus, $C_4 \sim 4$. Again applying the fact that arithmetic mean is at least as much as the geometric mean, we have

$$
\int_0^\infty P_2(x)dx \le \inf_{x>0}\{x + C_2 n^{-2}x^{-3}\} = 2C_2^{1/4}n^{-1/2}, \qquad (7.16)
$$

$$
\text{where, } C_2^{1/4} = \frac{2|b\bar{b}|}{\underline{r}}\left(\frac{4C_4}{3}\right)^{1/4}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i-\eta)^2}.
$$

Finally, combining (7.14) and (7.16) with (7.12) yields

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\tau\ge 0}\left|n^{-1}\log\left\{1-c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right|\right] \le 2\big(C_1^{1/2}+C_2^{1/4}\big)n^{-1/2} \\
&\le c_0|b\bar{b}|\underline{r}^{-1}\max\left\{1, \left\{\frac{1}{n}\sum_{i=1}^{n}(\theta_i-\eta)^2\right\}^{1/2}\right\}n^{-1/2}
\end{aligned}
$$

where $c_0 = 4(\sqrt{2}+(4\,C_4/3)^{1/4}) \approx 11.74$. This completes the proof for $\alpha \ne \pm 1$. The proof for $\alpha = \pm 1$ follows the same line as in the above.

7.5 Proof for Theorem 4.2

Let $\varepsilon_i := (X_i - \theta_i)$ for $i = 1,\ldots,n$. Let $\omega(\tau) := \tau/(\tau+1)$ and $\bar{\omega}(\tau) := 1-\omega(\tau)$. Also let $m_i(\tau) := \bar{\omega}_i^2(\tau)/(r_i + b^2\omega(\tau) + b\bar{b}\omega^2(\tau))$ for $i = 1,\ldots,n$. Consider the case with $\alpha \ne \pm 1$. Start with the following representation of the target quantity:

$$
\mathbb{E}\left[\sup_{\tau\ge 0, |\eta|\le\hat{q}}\left|n^{-1}\log\left\{1-c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\right\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\right|\right] \le \int_0^\infty \{P_1(x)+P_2(x)+P_3(x)\}dx,
$$

$$
\text{where, } P_1(x) = \Pr\left(\sup_{\tau\ge 0}\left|\sum_{i=1}^{n}(\varepsilon_i^2-1)m_i(\tau)\right| > \frac{nx}{2|b\bar{b}|}\right),
$$

$$
P_2(x) = \Pr\left(\sup_{\tau\ge 0}\left|\sum_{i=1}^{n}\theta_i\varepsilon_i m_i(\tau)\right| > \frac{nx}{2|b\bar{b}|}\right) \text{ and,}
$$

$$P_3(x) = \Pr\left(\sup_{\tau\ge 0, |\eta|\le \max\{\hat{q}_1, \hat{q}_2\}} \left|\sum_{i=1}^n \eta\varepsilon_i m_i(\tau)\right| > \frac{nx}{2|b\bar{b}|}\right).$$

First consider bounding $P_1(x)$ and $P_2(x)$. From (7.14) and (7.16), we get

$$\int_0^\infty P_1(x)dx \le 2C_1^{1/2}n^{-1/2} \text{ and } \int_0^\infty P_2(x) \le 2C_2^{1/2}n^{-1/2} \tag{7.17}$$

with constants C_1 and C_2 given in (7.14) and (7.16). Next consider bounding $P_3(x)$. Similarly as in (7.13), we get

$$\sup_{\tau\ge 0, |\eta|\le \max\{\hat{q}_1, \hat{q}_2\}} \left|\sum_{i=1}^n \eta\varepsilon_i m_i(\tau)\right| \le \max\{|\hat{q}_1|, |\hat{q}_2|\}\underline{r}^{-1} \sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|,$$

where $\underline{r} := \inf_i r_i$. This yields

$$\int_0^\infty P_3(x)dx \le \frac{2|b\bar{b}|}{n}\underline{r}^{-1}\mathbb{E}\left[\max\{|\hat{q}_1|, |\hat{q}_2|\}\sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|\right].$$

Here, the Markov inequality gives

$$\max\{|\hat{q}_1|, |\hat{q}_2|\} \le c_q\frac{\sum_{i=1}^n |X_i|}{n} \text{ with } c_q := \max\left\{\frac{2}{q}, \frac{1}{1-q/2}\right\},$$

and then this gives

$$\int_0^\infty P_3(x)dx \le \frac{2c_q|b\bar{b}|}{\underline{r}}\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[|X_i|\sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|\right] \le \frac{2c_q|b\bar{b}|}{\underline{r}}(S_1+S_2),$$

$$\text{where, } S_1 := \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[|\varepsilon_i|\sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|\right] \text{ and, } S_2 := \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[|\theta_i|\sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|\right].$$

Consider bounding S_1. Applying the Cauchy–Schwarz and the Lévy inequalities to S_1 yields

$$S_1 \le \frac{1}{n}\left\{\mathbb{E}\left[\sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|^2\right]\right\}^{1/2} \le \frac{\sqrt{2}}{n}\left\{\mathbb{E}\left[\left|\sum_{k=1}^n \varepsilon_k\right|^2\right]\right\}^{1/2} \le \sqrt{2}n^{-1/2}.$$

Consider bounding S_2. Applying the Cauchy–Schwarz, the Lévy, and again the Cauchy–Schwarz inequalities to S_2 yields

$$\begin{aligned} S_2 &\le \frac{1}{n}\left\{\frac{\sum_{i=1}^n \theta_i^2}{n}\right\}^{1/2}\mathbb{E}\left[\sup_{1\le j\le n}\left|\sum_{k=j}^n \varepsilon_k\right|\right] \le \frac{2}{n}\left\{\frac{\sum_{i=1}^n \theta_i^2}{n}\right\}^{1/2}\mathbb{E}\left[\left|\sum_{k=1}^n \varepsilon_k\right|\right] \\ &\le 2\left\{\frac{\sum_{i=1}^n \theta_i^2}{n}\right\}^{1/2} n^{-1/2}. \end{aligned}$$

Therefore, we obtain

$$\int_0^\infty P_3(x)dx \le 2C_3^{1/2}n^{-1/2} \text{ with } C_3^{1/2} := \frac{2c_q|b\bar{b}|}{\underline{r}}\Big\{1 + \Big(\frac{\sum_{i=1}^n \theta_i^2}{n}\Big)^{1/2}\Big\}. \tag{7.18}$$

Finally, by combining (7.17) with (7.18), we get

$$\mathbb{E}\Big[\sup_{\tau\ge 0, |\eta|\le \hat{q}} \big|n^{-1}\log\{1 - c_\alpha\, R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\big|\Big] \le 2(C_1^{1/2} + C_2^{1/2} + C_3^{1/2})n^{-1/2}$$

which completes the proof of the theorem.

7.6 Proof for Theorem 4.3

We prove the statement of the theorem for $\gamma = \gamma(\alpha) := \operatorname{sign}(\alpha^2 - 1) = -1$. The other case follows similarly. Consider an evaluation of the risk function at our proposed hyper-parameter values:

$$\check{R}_{n,\alpha}(\boldsymbol{\theta};\boldsymbol{X}) = R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\big|_{\eta=\hat{\eta}_{n,\alpha},\tau=\hat{\tau}_{n,\alpha}}.$$

Note that $\check{R}_{n,\alpha}$ is not a risk function as it depends on the data $\boldsymbol{X}$ through $\hat{\eta}_{n,\alpha}$ and $\hat{\tau}_{n,\alpha}$. As such, $\check{R}_{n,\alpha}(\boldsymbol{\theta};\boldsymbol{X}) \ne R_{n,\alpha}(\boldsymbol{\theta};\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha}) = \mathbb{E}\, L_{n,\alpha}(\boldsymbol{\theta},\hat{p}[\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha}])$. Consider

$$\mathcal{E}_{n,\alpha}(\boldsymbol{\theta}) := \mathbb{E}\big(n^{-1}\log(1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta};\eta^{\textbf{or}}_{n,\alpha},\tau^{\textbf{or}}_{n,\alpha})) - n^{-1}\log(1 - c_\alpha \check{R}_{n,\alpha}(\boldsymbol{\theta};\boldsymbol{X}))\big) \tag{7.19}$$

As $n \to \infty$, we provide an upper bound on $\mathcal{E}_{n,\alpha}(\boldsymbol{\theta})$. From Theorem 4.1, we have

$$\mathbb{E}\Big[\sup_{\tau\ge 0, \eta\in[\hat{q}_1,\hat{q}_2]} \Big|n^{-1}\log(1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)) - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\Big|\Big] \le 2(C_1^{1/2} + C_2^{1/2} + C_3)n^{-1/2}$$

with constants C_1, C_2, C_3 given in (7.14), (7.16), (7.18). Also, from the definition of $(\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha})$, we know that

$$\hat{\mathcal{H}}_n(\eta^{\textbf{or}}_{n,\alpha},\tau^{\textbf{or}}_{n,\alpha};\alpha) \le \hat{\mathcal{H}}_n(\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha};\alpha).$$

Combining them, we get

$$\begin{aligned}\mathcal{E}_{n,\alpha}(\boldsymbol{\theta}) &= \mathbb{E}\Big[n^{-1}\log(1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta};\eta^{\textbf{or}}_{n,\alpha},\tau^{\textbf{or}}_{n,\alpha})) - n^{-1}\log(1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta};\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha}))\Big]\\ &\le \mathbb{E}\Big[\hat{\mathcal{H}}_n(\eta^{\textbf{or}}_{n,\alpha},\tau^{\textbf{or}}_{n,\alpha};\alpha) - \hat{\mathcal{H}}_n(\hat{\eta}_{n,\alpha},\hat{\tau}_{n,\alpha};\alpha)\Big]\\ &\quad + 2\mathbb{E}\Big[\sup_{\tau\ge 0, \eta\in[\hat{q}_1,\hat{q}_2]} \big|n^{-1}\log\{1 - c_\alpha R_{n,\alpha}(\boldsymbol{\theta};\eta,\tau)\} - \hat{\mathcal{H}}_n(\eta,\tau;\alpha)\big|\Big]\\ &\le 4(C_1^{1/2} + C_2^{1/2} + C_3)n^{-1/2}.\end{aligned}$$

This shows $\limsup_{n\to\infty} n^{1/2}\rho_{n,\alpha}(\boldsymbol{\theta}) < \infty$. Again, by assumption **[A3]**, for any $\varepsilon > 0$ the event $\{|\tau^{\mathsf{or}} - \hat{\tau}_{n,\alpha}| > \varepsilon,\, |\eta^{\mathsf{or}}_{n,\alpha} - \hat{\eta}_{n,\alpha}| > \varepsilon\}$ is contained in the event $\{\sqrt{n}\mathcal{E}_{n,\alpha}(\boldsymbol{\theta}) \ge C\}$ with $C > 4(C_1^{1/2} + C_2^{1/2} + C_3)$. Thus, $\Pr(|\tau^{\mathsf{or}}_{n,\alpha} - \hat{\tau}_{n,\alpha}| > \varepsilon,\, |\eta^{\mathsf{or}}_{n,\alpha} - \hat{\eta}_{n,\alpha}| > \varepsilon) \to 0$ as $n \to \infty$.

Supplementary Material

Supplementary materials for Optimal Shrinkage Estimation of Predictive Densities under α–divergences. The supplement contains two numerical examples regarding the prdes discussed in this paper.

References

[1] Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge University Press.

[2] Amari, S.-I. (2009). "α-Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes." *IEEE Transactions on Information Theory*, 55(11): 4925–4931.

[3] Barron, A., Rissanen, J., and Yu, B. (1998). "The minimum description length principle in coding and modeling." *IEEE Transactions on Information Theory*, 44(6): 2743–2760.

[4] Barron, A. R., Gyorfi, L., and van der Meulen, E. C. (1992). "Distribution estimation consistent in total variation and in two types of information divergence." *IEEE transactions on Information Theory*, 38(5): 1437–1454.

[5] Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distributions." *Bulletin of the Calcutta Mathematical Society*, 99–109.

[6] Brown, L. D., George, E. I., and Xu, X. (2008). "Admissible predictive density estimation." *Ann. Statist.*, 36(3): 1156–1170.

[7] Cichocki, A. and Amari, S.-i. (2010). "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities." *Entropy*, 12(6): 1532–1568.

[8] Clarke, B. and Yuan, A. (2010). "Reference priors for empirical likelihoods." *Frontiers of Statistical Decision Making and Bayesian Analysis: In honor of James O. Berger*, 56–68.

[9] Corcuera, J. and Giummolè, F. (1999). "A Generalized Bayes Rule for Prediction." *Scandinavian Journal of Statistics*, 26(2): 265–279.

[10] Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

[11] Efron, B. and Morris, C. (1973). "Stein's estimation rule and its competitors—an empirical Bayes approach." *Journal of the American Statistical Association*, 68(341): 117–130.

[12] Fourdrinier, D., Marchand, É., Righi, A., and Strawderman, W. E. (2011). "On improved predictive density estimation with parametric constraints." *Electron. J. Stat.*, 5: 172–191.

[13] Fourdrinier, D., Strawderman, W. E., and Wells, M. T. (2018). *Shrinkage estimation*. Springer.

[14] Geisser, S. (1993). *Predictive inference*, volume 55 of *Monographs on Statistics and Applied Probability*. New York: Chapman and Hall. An introduction.

[15] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

[16] Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6): 997–1016.

[17] George, E. I., Liang, F., and Xu, X. (2006). "Improved minimax predictive densities under Kullback-Leibler loss." *Ann. Statist.*, 34(1): 78–91.

[18] — (2012). "From minimax shrinkage estimation to minimax shrinkage prediction." *Statist. Sci.*, 27(1): 82–94.

[19] Ghosh, M. and Kubokawa, T. (2018). "Hierarchical Bayes versus empirical Bayes density predictors under general divergence loss." *Biometrika*.

[20] Ghosh, M., Kubokawa, T., and Datta, G. S. (2019). "Density Prediction and the Stein Phenomenon." *Sankhya A*, 1–23.

[21] Ghosh, M., Mergel, V., and Datta, G. S. (2008). "Estimation, prediction and the Stein phenomenon under divergence loss." *J. Multivariate Anal.*, 99(9): 1941–1961.

[22] Ghosh, M., Mergel, V., and Liu, R. (2011). "A general divergence criterion for prior selection." *Annals of the Institute of Statistical Mathematics*, 63(1): 43–58.

[23] Gül, G. and Zoubir, A. M. (2016). "Robust hypothesis testing with \alpha-divergence." *IEEE Transactions on Signal Processing*, 64(18): 4737–4750.

[24] Hernández-Lobato, J. M., Li, Y., Rowland, M., Hernández-Lobato, D., Bui, T., and Turner, R. (2016). "Black-box α-divergence minimization." In *Proceedings of The 33rd International Conference on Machine Learning*. International Machine Learning Society.

[25] Ibragimov, R. and Sharakhmetov, S. (2002). "The exact constant in the Rosenthal inequality for random variables with mean zero." *Theory of Probability & Its Applications*, 46(1): 127–132.

[26] James, W. and Stein, C. M. (1961). "Estimation With Quadratic Loss." In *Proceedings of the 4th Berkeley Symposium on Probability and Statistics*, 367–379.

[27] Kato, K. (2009). "Improved prediction for a multivariate normal distribution with unknown mean and variance." *Ann. Inst. Statist. Math.*, 61(3): 531–542.

[28] Kempthorne, P. J. et al. (1988). "Controlling risks under different loss functions: The compromise decision problem." *Annals of statistics*, 16(4): 1594–1608.

[29] Komaki, F. (2001). "A shrinkage predictive distribution for multivariate normal observables." *Biometrika*, 88(3): 859–864.

[30] — (2004). “Simultaneous prediction of independent Poisson observables.” *Ann. Statist.*, 32(4): 1744–1769.

[31] Kubokawa, T., Marchand, É., Strawderman, W. E., and Turcotte, J.-P. (2013). “Minimaxity in predictive density estimation with parametric constraints.” *Journal of Multivariate Analysis*, 116: 382–397.

[32] Liang, F. (2002). *Exact minimax procedures for predictive density estimation and data compression*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Yale University.

[33] Liang, F. and Barron, A. (2005). *Exact Minimax Predictive Density Estimation and MDL*, chapter 7, 177–194. Advances in Minimum Description Length: Theory and Applications (P. Grunwald, I. Myung and M. Pitt eds). MIT Press.

[34] Liese, F. and Vajda, I. (2006). “On divergences and informations in statistics and information theory.” *IEEE Transactions on Information Theory*, 52(10): 4394–4412.

[35] L’Moudden, A. and Marchand, É. (2018). “On Predictive Density Estimation under α-divergence Loss.” *arXiv preprint arXiv:1806.02600*.

[36] Maruyama, Y., Matsuda, T., and Ohnishi, T. (2019). “Harmonic Bayesian prediction under α-divergence.” *IEEE Transactions of Information Theory*.

[37] Maruyama, Y. and Strawderman, W. (2010). “Bayesian predictive densities for linear regression models under α-divergence loss: some results and open problems.” Manuscript available at :"http://arxiv.org/abs/1002.3786v1".

[38] Mukherjee, G. (2013). “Sparsity and Shrinkage in Predictive Density Estimation.” Ph.D. thesis, Stanford University.

[39] Mukherjee, G. and Johnstone, I. M. (2015). “Exact minimax estimation of the predictive density in sparse Gaussian models.” *Annals of Statistics*.

[40] — (2017). “On Minimax Optimality of Sparse Bayes Predictive Density Estimates.” *arXiv preprint arXiv:1707.04380*.

[41] Nguyen, X., Wainwright, M. J., Jordan, M. I., et al. (2009). “On surrogate loss functions and f-divergences.” *The Annals of Statistics*, 37(2): 876–904.

[42] Pardoe, I. (2001). “A Bayesian sampling approach to regression model checking.” *Journal of Computational and Graphical Statistics*, 10(4): 617–627.

[43] Robbins, H. (1964). “The empirical Bayes approach to statistical decision problems.” *The Annals of Mathematical Statistics*, 35(1): 1–20.

[44] Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M., and Verri, A. (2004). “Are loss functions all the same?” *Neural computation*, 16(5): 1063–1076.

[45] Suzuki, T. and Komaki, F. (2010). “On Prior Selection and Covariate Shift of β-Bayesian prediction under α-divergence risk.” *Comm. Statist. Theory and Methods*, 39: 1655–1673.

[46] Sweeting, T. J., Datta, G. S., and Ghosh, M. (2006). "Nonsubjective priors via predictive relative entropy regret." *The Annals of Statistics*, 441–468.

[47] Tan, Z. (2015). "Improved minimax estimation of a multivariate normal mean under heteroscedasticity." *Bernoulli*, 21(1): 574–603.

[48] Tay, A. S. and Wallis, K. F. (2000). "Density forecasting: a survey." *Journal of forecasting*, 19(4): 235–254.

[49] Taylor, J. W. and Buizza, R. (2004). "A comparison of temperature density forecasts from GARCH and atmospheric models." *Journal of Forecasting*, 23(5).

[50] Wang, D., Liu, H., and Liu, Q. (2018). "Variational inference with tail-adaptive f-divergence." In *Advances in Neural Information Processing Systems*, 5737–5747.

[51] Weinstein, A., Ma, Z., Brown, L. D., and Zhang, C.-H. (2018). "Group-linear empirical Bayes estimates for a heteroscedastic normal mean." *Journal of the American Statistical Association*, 113(522): 698–710.

[52] Xie, X., Kou, S., and Brown, L. (2016). "Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance." *Annals of statistics*, 44(2): 564.

[53] Xie, X., Kou, S., and Brown, L. D. (2012). "SURE estimates for a heteroscedastic hierarchical model." *Journal of the American Statistical Association*, 107(500).

[54] Xu, X. (2005). "Estimation of high dimensional predictive densities." Ph.D. thesis, University of Pennsylvania.

[55] Xu, X. and Liang, F. (2010). "Asymptotic minimax risk of predictive density estimation for non-parametric regression." *Bernoulli*, 16(2): 543–560.

[56] Xu, X. and Zhou, D. (2011). "Empirical Bayes predictive densities for high-dimensional normal models." *J. Multivariate Analysis*, 102(10): 1417–1428.

[57] Yano, K., Kaneko, R., and Komaki, F. (2021). "Minimax Predictive Density for Sparse Count Data." Forthcoming in Bernoulli.

[58] Yano, K. and Komaki, F. (2017). "Asymptotically minimax prediction in infinite sequence models." *Electronic Journal of Statistics*, 11(2): 3165–3195.

[59] Yuan, A. and Clarke, B. (1999). "An information criterion for likelihood selection." *IEEE Transactions on Information Theory*, 45(2): 562–571.

[60] Zhang, C.-H. (2003). "Compound decision theory and empirical Bayes methods: invited paper." *Ann. Statist.*, 31(2): 379–390.

[61] Zhang, F., Shi, Y., Ng, H. K. T., and Wang, R. (2017). "Information Geometry of Generalized Bayesian Prediction Using α-Divergences as Loss Functions." *IEEE Transactions on Information Theory*, 64(3): 1812–1824.

Acknowledgments

The research was partly supported by NSF grants DMS-1811866, DMS-1916245 and JST grant KAKENHI 19K20222.

Bayesian Analysis , Number , pp.

Supplementary materials for "Optimal Shrinkage Estimation of Predictive Densities under α–divergences"

Edward George[*], Gourab Mukherjee[†] and Keisuke Yano[‡]

1 Two Examples of Predictive Usage

We describe two examples to motivate the usage of our proposed prdes. First, consider predicting the mean cellular expression patterns of sup-populations of infected cells in $l = 1, \dots, L$ replicated experiments. In experiment l for protein i, we observe the expression (in arcsinh scale) $v_{ij}^{(l)}$ for $j = 1, \dots, k_v^{(l)}$ cells. Consider a Gaussian additive model for the protein intensity recorded from each cell:

$$v_{ij}^{(l)} = \theta_{i,v} + \sigma_{i,v}^{(l)} \epsilon_{ij}^{(l)},\ j = 1, \dots, k_v^{(l)},$$

where, $\theta_{i,v}$ is the average expression level and $\epsilon_{ij}^{(l)}$ are i.i.d. from standard normal distribution. In these virology experiments the percentage of infection is very low [1] and typically $k_v^{(l)}$ in the order of hundreds. As virus entry in cells and subsequent infection is an endogenous process with experimenter having no control, $k_v^{(l)}$ vary across replications l. Our object of interest is the average observed expression vector $\bar{\boldsymbol{v}}^{(l)}$ which has the same mean but reduced variance:

$$\bar{v}_i^{(l)} = \frac{1}{k_v^{(l)}} \sum_{j=1}^{k_v^{(l)}} v_{ij}^{(l)} \overset{d}{=} \theta_{i,v} + \bar{\sigma}_i^{(l)} Z_i^{(l)} \text{ where } \bar{\sigma}_i^{(l)} = \frac{\sigma_{i,v}^{(l)}}{\sqrt{k_v^{(l)}}} \text{ for } i = 1, \dots, 24 \text{ and } l = 1, \dots, L,$$

where, $Z_i^{(l)}$ are i.i.d. from standard normal. The cellular variability $\sigma_{i,v}^2$ can be well estimated based on the variability of other types of infected cells from those experiments. Based on observing $\bar{\boldsymbol{v}} = \bar{\boldsymbol{v}}^{(1)}$ and knowing $\{k_v^{(l)} : l = 1, \dots, L\}$ we considering predicting the average expression levels $\bar{\boldsymbol{v}}^{(l)}$ for $l > 1$. Here, $L = 50$. We consider BH and KL losses based prdes $\hat{p}_{\mathsf{H}}$, $\hat{p}_{\mathsf{KL}}$ in $\mathcal{S}$ that are tuned by our proposed method as well as the completely non-informative prior based prde $\hat{p}_{\mathsf{NI}}$. We also consider $\hat{p}_{\mathsf{Homo}}$ the prde in $\mathcal{S}$ that is tuned ignoring the heteroskedasticity in the data as well as two plugin prdes (a) $\hat{p}_{\mathsf{P}}^{\mathsf{C}}$ centered on the canonical mean estimator $\bar{\boldsymbol{v}}$ (b) $\hat{p}_{\mathsf{P}}^{\mathsf{JS}}$ centered on the extended James-Stein (JS) estimator

$$\hat{\theta}_i^{\mathsf{JS}}(\bar{\boldsymbol{v}}) = \hat{\eta}_{\mathsf{JS}} + \left(1 - \frac{n-3}{\sum_{i=1}^n \bar{\sigma}_i^{-2} (\bar{v}_i - \hat{\eta}_{\mathsf{JS}})^2}\right)_+ (\bar{v}_i - \hat{\eta}_{\mathsf{JS}}) \text{ with } \hat{\eta}_{\mathsf{JS}} = \frac{\sum_{i=1}^n \bar{\sigma}_i^{-2} \bar{v}_i}{\sum_{i=1}^n \bar{\sigma}_i^{-2}}\ .$$

[*] Department of Statistics, University of Pennsylvania, edgeorge@wharton.upenn.edu
[†] Department of Data Sciences and Operations, University of Southern California, gourab@usc.edu
[‡] The Institute of Statistical Mathematics, yano@ism.ac.jp

Note, that for these prediction problem $r_i^{(l)} = (\bar{\sigma}_i^{(l)}/\bar{\sigma}_i^{(1)})^2$ vary greatly across i and l. In Table 1 we report the average (across l) volumes of the 90% and 95% prediction intervals (PIs) as well as the average number of cases $\bar{v}_i^{(l)}$ falls outside the interval which is reported as non-coverage of the intervals. It was witnessed that though the plugin prdes produced prediction regions with the minimal volume, they are not conservative. All the other prdes produced conservative prediction intervals with $\hat{p}_{\mathsf{H}}$ having substantially lower volumes than the others.

Table 1: We report the average non-coverage and the average volume of 90% and 95% prediction intervals that are constructed using different prdes of the infected mean expression levels.

PRDE	**95% Prediction Interval**		**90% Prediction Interval**	
	Non-coverage	**Volume**	**Non-coverage**	**Volume**
BH	5.09%	30.745	7.87%	28.203
Kullback Leibler	2.31%	35.595	2.31%	32.652
Homoskedastic	3.70%	36.024	4.63%	33.046
Non-Informative	3.70%	36.108	4.63%	33.122
Plugin Canonical	10.65%	24.789	13.89%	22.740
Plugin JS	10.65%	24.789	14.81%	22.740

Next, we consider a different problem where the goal is finding significant contrasts in expressions between infected and uninfected population. In virology it is important to predict the difference in the average protein expression between the infected and uninfected cell populations [2]. For uninfected cells we observe $u_{ij}^{(l)}$ – the expression of the ith protein across $j = 1, \ldots, k_u^{(l)}$ cells from $l = 1, \ldots, L$ samples. Here, $L = 20$. The mean expression profile $\boldsymbol{\theta}_u = \{\theta_{i,u} : 1 \le i \le n\}$ in uninfected cells is possibly different form the mean profile $\boldsymbol{\theta}_v^{(l)}$. We model

$$u_{ij}^{(l)} = \theta_{i,u}^{(l)} + \sigma_{i,u}^{(l)}\,\epsilon_{ijv}^{(l)} \text{ and } v_{i\tilde{j}}^{(l)} = \theta_{i,v}^{(l)} + \sigma_{i,v}^{(l)}\,\epsilon_{i\tilde{j}v}^{(l)}\ ,$$

where, $\epsilon_{iju}^{(l)}$ and $\epsilon_{i\tilde{j}v}^{(l)}$ are independent white noise and $j = 1, \ldots, k_u^{(l)}$ and $\tilde{j} = 1, \ldots, k_v^{(l)}$. The average expression difference vector $\boldsymbol{d} = \{\bar{v}_i^{(l)} - \bar{u}_i^{(l)} : 1 \le i \le 24\}$ follows $d_i^{(l)} = \theta_{i,d}^{(l)} + \sigma_{i,d}^{(l)}\, Z_i^{(l)}$ where $\theta_{i,d}^{(l)} = \theta_{i,v}^{(l)} - \theta_{i,u}^{(l)}$ and $\sigma_{i,d}^{(l)} = \{(\sigma_{i,v}^{(l)})^2/k_v^{(l)} + (\sigma_{i,u}^{(l)})^2/k_u^{(l)}\}^{1/2}$. The mean difference between infected and uninfected expressions corresponds to fold changes in proteins due to viral infection. For $L = 20$ samples, we observe pre-infection and post VZV infection protein expressions for $i = 1, 2, \ldots, 24$ proteins over the baseline. These proteins can be classified into two functional types (a) surface (b) signaling. There is discernible difference in the variability and mean expression patterns between these two groups. Figure 1 (left panel) show the variance of the 11 surface proteins (in red) and the 13 signaling proteins (in blue).

For $l = 1, \ldots, L$, we construct 90% and 95% simultaneous PIs for the mean expression difference between infected and uninfected cells. Figure 1 (right panel) shows the

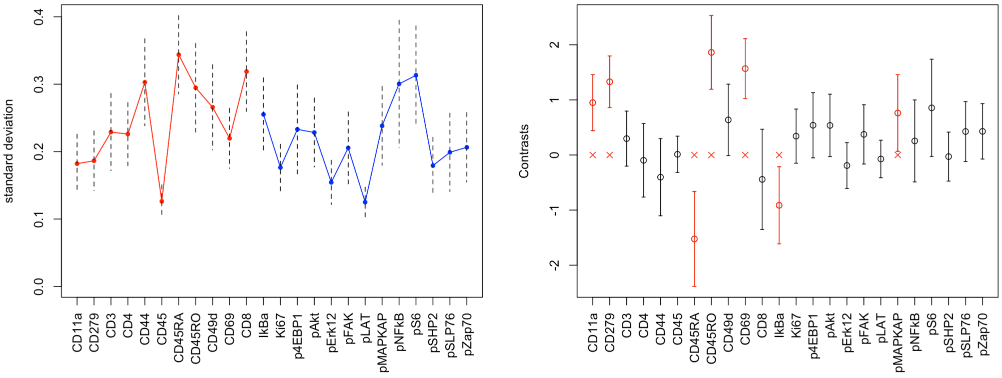

Figure 1: Left panel: Standard deviations (SD) of the VZV infected average expressions for 11 surface proteins (in red) and the 13 signaling proteins (in blue). The average SD across 20 samples and their one SD ranges is shown. Right panel: 95 % Predictive intervals based on $\hat{p}_{\mathsf{H}}$ for the difference of means between virus and uninfected sample. In red corresponds to markers where there is significant difference with the x denoting that 0 is outside the 95% interval.

Table 2: We report the number of proteins (# Sig) with significantly different expressions between infected and uninfected samples for the 95% and 90% prediction intervals. The reproducibility (R-index) of these significant contrasts across 19 experiments and the average volumes of the PIs are also reported.

PRDE	**95% Prediction Interval**			**90% Prediction Interval**		
	# Sig	**R-index**	**Volume**	**# Sig**	**R-index**	**Volume**
BH	6.6	93.33%	38.478	7.4	92.50%	35.296
Kullback Leibler	5.3	94.58%	44.204	6.2	93.33%	40.549
Homoskedastic	5.4	94.17%	44.680	6.5	93.33%	40.986
Non-Informative	5.3	93.33%	44.874	6.6	93.33%	41.164
Plugin Canonical	9.1	88.96%	31.731	9.9	86.88%	29.107
Plugin JS	9.1	88.33%	31.731	10	86.25%	29.107

95% PI from $\hat{p}_{\mathsf{H}}$ in $l = 1$. The intervals marked in red do not contain the origin. So, at 5% level of significance they denote the proteins which has different expression in the infected population. Up regulation of only one signaling protein was witnessed in the infected sample along with changes in six cell surface proteins (two down and four up regulation). It shows down regulation of naive marker (CD45RA) accompanied by corresponding up regulation of the memory marker (CD45RO) which was reported in [3]. In Table 2, we report the average number of significant contrasts based on 90% and 95% prediction intervals. For $c = 0.05$ and 0.1 we find significant contrasts between infected and uninfected samples for $l = 1, \ldots, L$ at level c based on $100(1 - c)$% PIs constructed using different prdes. The contrasts $\theta_{i,d}^{(l)}$ is expected to be invariant across $l = 1, \ldots, L$. In table 2 we report the reproducibility of these findings across the different samples by the reproducibility index (R-index). We define R-index as $24^{-1} \sum_{i=1}^{24} d_i \vee (1 - d_i)$ where $d_i = L^{-1} \sum_{l=1}^{L} I\{\theta_{i,d}^{(l)} \neq 0\}$.

Prdes that produce PIs with higher R-index and lower average volumes are desired. Plugin prdes do not attain desired R-index for both cases $c = 0.05$ and 0.1. The other prdes have reasonable R-indices. Among those the BH loss based prde in $\mathcal{S}$ produces PIs with the least average volume and also with the maximum average discoveries of significant contrasts. However, note that here we do not have any criterion or recommendation for prescribing what would be a good α-divergence metric for producing prediction intervals or for hypothesis testing. In real-world applications, we need decision-theoretic formulation of a loss function that is tailored to the specific inferential task at hand. We need a suitable loss function (in this case a good choice of α) that is carefully designed based on the application at hand and without looking at the data. Here, we do not conduct any introspection on what would be a good choice of α but present the two examples merely for illustrating results based on different prdes used in the paper. We witness sub-optimal performance of the aforementioned plug-in prde but can not suggest any optimal choice from the other prdes.

References

[1] Banerjee, T., Bhattacharya, B. B., and Mukherjee, G. (2020). "A Nearest-Neighbor Based Nonparametric Test for Viral Remodeling in Heterogeneous Single-Cell Proteomic Data." *Annals of Applied Statistics*.

[2] Sen, A., Rothenberg, M. E., Mukherjee, G., Feng, N., Kalisky, T., Nair, N., Johnstone, I. M., Clarke, M. F., and Greenberg, H. B. (2012). "Innate immune response to homologous rotavirus infection in the small intestinal villous epithelium at single-cell resolution." *Proceedings of the National Academy of Sciences*, 109(50): 20667–20672.

[3] Sen, N., Mukherjee, G., Sen, A., Bendall, S. C., Sung, P., Nolan, G. P., and Arvin, A. M. (2014). "Single-cell mass cytometry analysis of human tonsil T cell remodeling by varicella zoster virus." *Cell reports*, 8(2): 633–645.